ED 032 214 SE 007 464

Final Report to the National Science Foundation on Contract NSF-C414 Task I.
American Chemical Society, Columbus, Ohio. Chemical Abstracts Service.
Spons Agency-National Science Foundation, Washington, D.C.
Pub Date Mar 69
Note-194p.
EDRS Price MF-$0.75 HC-$9.80
Descriptors-*Chemistry, Information Dissemination, *Information Processing, *Information Science, Systems Development
Identifiers-Chemical Abstracts Service, Chemical Registry System

Chemical Abstracts Service developed and expanded a computer-based Chemical Registry System and operated it on a large-scale pilot basis. Some 1,744,319 registry transactions were made, resulting in the addition of 988,806 unique substances to the Registry Files. Continual effort has been made to improve computer capabilities and procedures, add new types of compounds, and improve user services and programs. In addition to registration, this included (1) the testing of alternative input methods for chemical information, including the development of special keyboards and keyboarding conventions; (2) the installation of several operational adjustments in the System; and (3) the development of computer support systems. Tabular data are included to show various operations and characteristics of the System. (RR)

F-NSF

ED032214

FINAL REPORT

to the

NATIONAL SCIENCE FOUNDATION

on

CONTRACT NSF-C414, TASK I

CHEMICAL ABSTRACTS SERVICE

AMERICAN CHEMICAL SOCIETY

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Columbus, Ohio March 1969

SE 007 464

FINAL REPORT

to the

NATIONAL SCIENCE FOUNDATION

on

CONTRACT NSF-C414, TASK I

CHEMICAL ABSTRACTS SERVICE

AMERICAN CHEMICAL SOCIETY

"PERMISSION TO REPRODUCE THIS
COPYRIGHTED MATERIAL HAS BEEN GRANTED
BY Fred A. Tate
Chemical Abstracts
TO ERIC AND ORGANIZATIONS OPERATING
UNDER AGREEMENTS WITH THE U.S. OFFICE OF
EDUCATION. FURTHER REPRODUCTION OUTSIDE
THE ERIC SYSTEM REQUIRES PERMISSION OF
THE COPYRIGHT OWNER."

Columbus, Ohio

March 1969

CONTENTS

Page

ABSTRACT. 1

I. INTRODUCTION. 2

II. OUTLINE OF REGISTRY OPERATIONS 5

III. NUMBERS AND SOURCES OF REGISTERED SUBSTANCES. 10

IV. REGISTRY INPUT METHODS. 16

- A. Input of Structural Information. 16
- B. Input of Nonstructural Information 22

V. IMPROVEMENTS AND EXTENSIONS IN THE REGISTRY 27

- A. Extensions of Registry Algorithm 27
- B. Computer-Checked Temporary Identification Numbers. 27
- C. "Dot-Disconnected" Convention. 28
- D. Extension of the Registry to New Classes of Compounds. 28
- E. Improved Text Descriptor Processing. 33

VI. DESKTOP ANALYSIS TOOLS 36

- A. Italicization. 37
- B. Capitalization 38
- C. Checking of Punctuation Consistency. 39
- D. Elimination of Invalid Characters. 39
- E. Printing of Diagnostic Comments. 39
- F. Nomenclature Sort Key Program. 41

VII. REDESIGN OF THE REGISTRY COMPUTER SYSTEMS AND REPROGRAMMING FOR THE IBM 360 COMPUTERS 46

- A. Reprogramming the Structure Registry 47
- B. Reprogramming of Nonstructural Systems 48

VIII. GLOSSARY 51

APPENDIXES

APPENDIX A

An Overview of the CAS Chemical Registry System

APPENDIX B

The Chemical Compound Registry

APPENDIX C

A Description of the CAS Chemical Registry System

APPENDIX D

The Generation of a Unique Machine Description for Chemical Structures -- A Technique Developed at Chemical Abstracts Service

APPENDIX E

The Computer-Based Subject Index Support System at Chemical Abstracts Service

APPENDIX F

Improvements in Structure Registry Effected with the Redesign and Reprogramming for IBM 360 Computers

APPENDIX G

Systems for Registering and Naming Polymers at Chemical Abstracts Service

LIST OF FIGURES

Page

REGISTRY SYSTEM CHART A 7

REGISTRY SYSTEM CHART B 8

REGISTRY SYSTEM CHART C 9

FIGURE 1: CAS Text-Typing Keyboard 18

FIGURE 2: CAS Structure-Typing Keyboard. 21

FIGURE 3: Example of a Computer-Produced Data Sheet. 24

FIGURE 4: Examples of Computer Produced Diagnostic Comments. 40

FIGURE 5: Illustration of Ordering on Compound Parent. 43

FIGURE 6: Illustration of Ordering Ignoring Prefixes 44

FIGURE 7: Illustration of Ordering by Numeric Value When Alphabetics Are Identical 45

FIGURE 8: Typed Structure For the B Chain of Bovine Insulin. 49

LIST OF TABLES

Page

TABLE I - Registry Status Summary Report. 12

TABLE 2 - Total Number of Registrations Performed through 16 March 1969 13

TABLE 3 - Summary of Registration 14

TABLE 4 - First Source of Names on Nomenclature File. 15

ABSTRACT

Under Contract NSF-C414 (1 June 1965 - 16 March 1969), Chemical Abstracts Service has developed and expanded a computer-based Chemical Registry System and operated it on a large-scale pilot basis. As a result of Task I, some 1,744,319 registry transactions were made, resulting in the addition of 988,806 unique substances to the Registry Files. Together with registrations from other sources, this brought the total Registry File to 1,079,551 unique substances.

Following the initial development of the Registry System, continual effort has been made to improve computer capabilities and procedures, add new types of compounds, and improve user services and programs. In addition to registration, this includes (1) the testing of alternative input methods for chemical information, including the development of special keyboards and keyboarding conventions that reduce the effort required to generate data in machine language for input to the computer; (2) the installation of several operational adjustments in the System that increase overall efficiency and broaden the range of compounds that are machine registered; and (3) the development of computer support systems that increase the productive efficiency of the chemical and clerical staff of the Chemical Registry System.

This Final Report describes the work performed under this contract and indicates the present status and operation of the Chemical Registry System. In addition, tabular data are included to show various operations and characteristics of the System.

I. INTRODUCTION

Between 1 June 1965 and 16 March 1969, Chemical Abstracts Service, a Division of the American Chemical Society, contracted with the National Science Foundation to undertake the experimental development and pilot operation of a computer-based Chemical Registry System. The function of this information-handling system is to organize and file information about chemical substances, identifying each substance on the basis of its conventional two-dimensional, structural diagram and resolving different versions of the same diagram so as to uniquely identify each substance. The Registry System files structural data, molecular formulas, names, and bibliographic citations in a set of interrelated manual and machine files, making possible the quick retrieval of various items of data about the substances filed.

The development of the Chemical Registry System was undertaken as a key step in building an operational, integrated, man-machine system for manipulating information about chemical substances -- a system that would be capable of high speed, flexibility, and depth in responding to the information needs of those who use chemical information. The system of registration provides a basis for organizing substance-oriented data selected from the full range of the scientific and technical literature. The Registry System was established on a computer basis to assure maximum consistency, efficiency, timeliness, and responsiveness of operation.

The value of a single, unified computer-based repository is readily apparent when it is noted that an estimated 85% of all published chemical literature relates in some way to chemical compounds or mixtures. Already, with just over four years of operation, the CAS Chemical Registry System includes information about more than one million substances, and eventually data on the estimated three million or more compounds reported in CA and Beilstein are expected to be incorporated into the System. Such a large-scale system, based on computer handling, allows correlations to be made between the items of data in the file which would not be practical with manual information resources. In addition, the data within the System can be retrieved in a variety of ways, and manipulated to make many types of subject-oriented, index-type listings available with a minimum of human intervention.

The Registry System operates by assigning a unique number, called a Registry Number, to each compound when it is first entered into the file. Whenever a compound which is already on file appears in a new reference, the previously assigned number is automatically recovered for use in filing the new reference. The System has three principal files of data: the atom-bond connection records of the chemical structure, the various forms of nomenclature associated with each substance, and bibliographic identification of the sources from which each compound was registered. A description of the Registry System's scope and potential use is provided in Appendixes A and B.

The objective of Task I of Contract NSF-C414 has been (1) the building up of the Registry Files by registering the compounds from a variety of sources including CAS internal reference files and the current literature as reported in CA; (2) accumulating performance data under large-scale pilot-plant conditions; (3) testing various methods of input to the system;

(4) broadening the range of compounds that can be handled in the machine system; and (5) developing and producing analysis tools and computer support systems with which the chemical and clerical staff of the Chemical Registry System can operate more efficiently.

II. OUTLINE OF REGISTRY OPERATIONS

Registry Charts A, B, and C (pages 7, 8, and 9), are included here to summarize the flow of data in the CAS Chemical Registry System for compounds registered as a result of processing information for Chemical Abstracts (CA) indexes. Appendix C describes Registry workflow in greater detail.

Registration is an identification procedure for chemical substances. By definition, registration is the process of determining whether or not a substance is in the Registry Files and of establishing a unique numerical label ("Registry Number") for each different registered substance. Thus, the Registry System depends upon the algorithmic manipulation of information in a conventional structural diagram to create one, and only one, machine record for each substance. This record must always be the same for each different substance -- it cannot vary with the size or orientation of the structural diagram, with arbitrary numbering conventions, or with varying chemical nomenclature. The algorithm used in the CAS Registry System is described in Appendix D, and has been mathematically proven to accomplish the unique identification of each substance. For each unique substance, a Registry Number is assigned as an identifying tag. This number is serially assigned by a computer algorithm; the numbers have no established pattern relating to the structural configuration of the compounds registered. Each Registry Number is a nine-digit number, the last digit being a machine-computed check digit for use in reducing the number of transcription errors.

The Registry Number functions as a machine address within the associated files, acting as a tie between the information related to a given compound in each of the System's files. For instance, it ties nomenclature and bibliographic information to the structural representation in the connection table file. It can also serve this function in files of conceptual information such as computer files of biochemical and physical property data, which may be developed by users of the System.

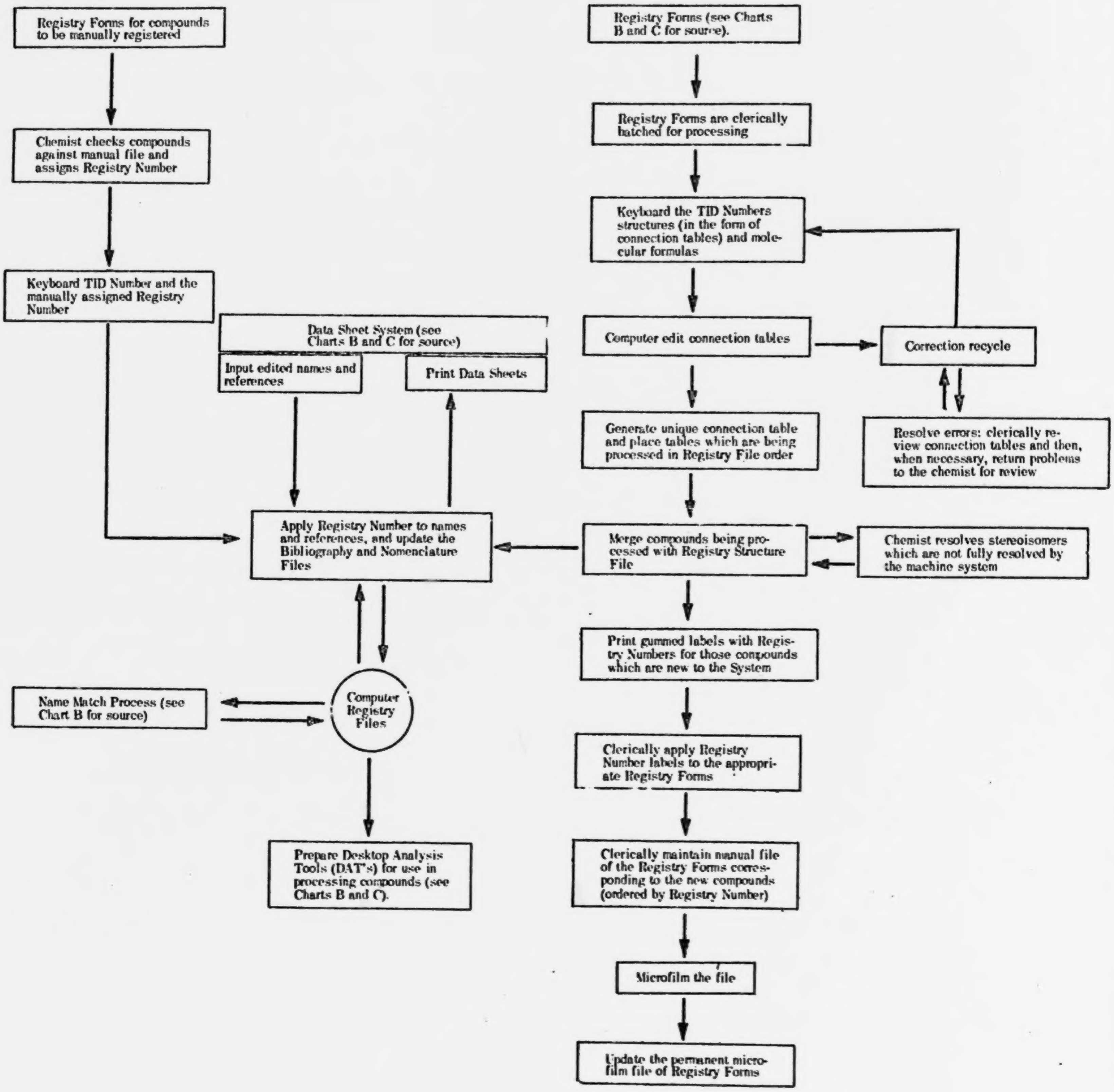
The Registry Number also has other uses such as providing a standardized, concise form of identification for organizing collections of substance-oriented data. Since the System provides unique identification for compounds, it eliminates the likelihood of filing information at two or more unrelated points under unrecognized synonymous names.

For certain types of substances -- a small minority -- structural definition is incomplete or conventions have not yet been established for registration. Such substances can be registered manually by a chemist working with a small set of files. The nonstructural information about manually registered compounds is added to the computer files and thus is available for use along with information on computer-registered substances.

REGISTRY SYSTEM CHART A

Data Flow in the CAS Compound Registry System

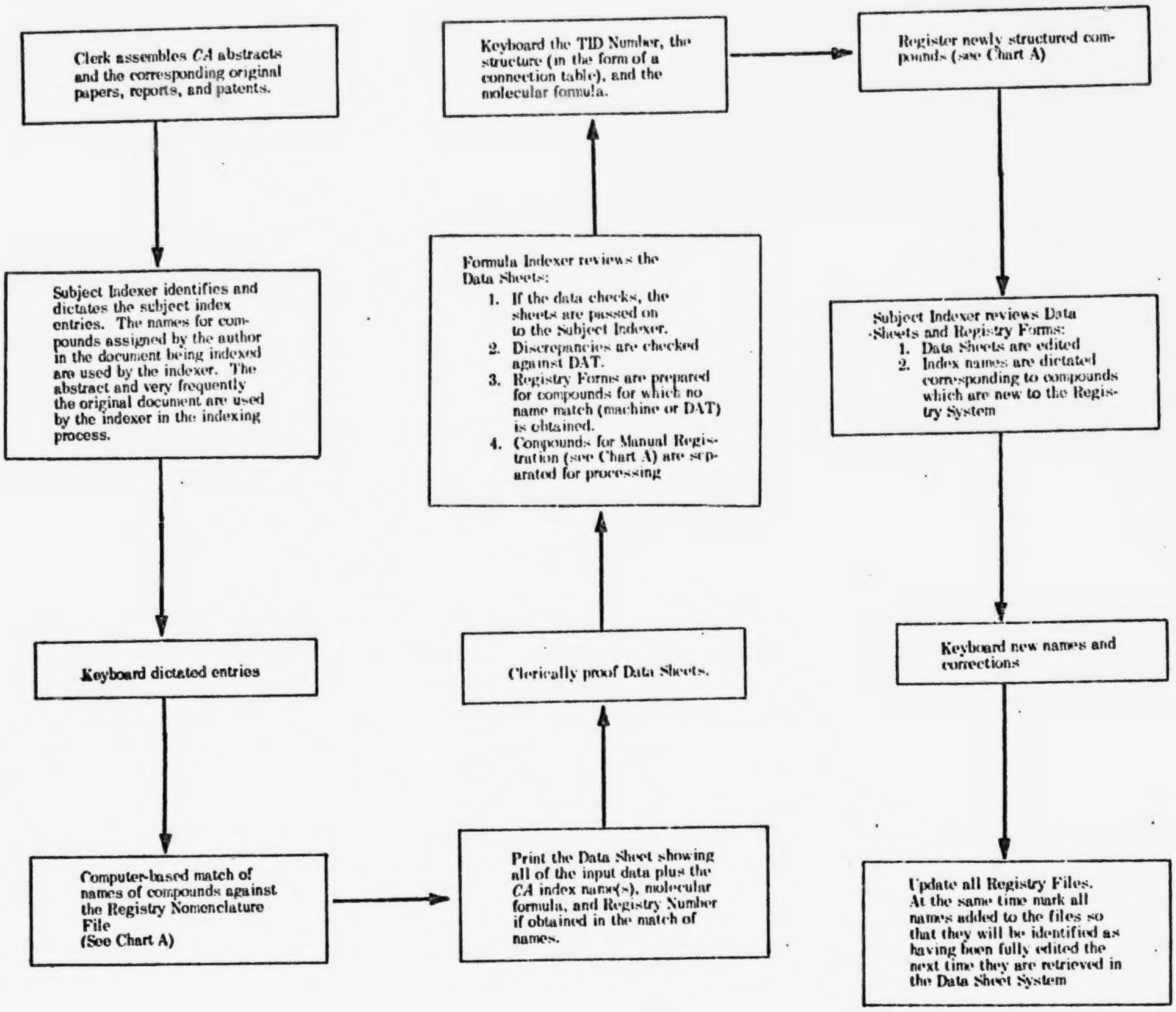
All steps are computer-based unless otherwise noted.


Registry Forms for compounds to be manually registered
Chemist checks compounds against manual file and assigns Registry Number
Keyboard TID Number and the manually assigned Registry Number
Data Sheet System (see Charts B and C for source)
Input edited names and references
Print Data Sheets
Apply Registry Number to names and references, and update the Bibliography and Nomenclature Files
Computer Registry Files
Name Match Process (see Chart B for source)
Prepare Desktop Analysis Tools (DAT's) for use in processing compounds (see Charts B and C).
Registry Forms (see Charts B and C for source).
Registry Forms are clerically batched for processing
Keyboard the TID Numbers structures (in the form of connection tables) and molecular formulas
Computer edit connection tables
Correction recycle
Resolve errors: clerically review connection tables and then, when necessary, return problems to the chemist for review
Generate unique connection table and place tables which are being processed in Registry File order
Merge compounds being processed with Registry Structure File
Chemist resolves stereoisomers which are not fully resolved by the machine system
Print gummed labels with Registry Numbers for those compounds which are new to the System
Clerically apply Registry Number labels to the appropriate Registry Forms
Clerically maintain manual file of the Registry Forms corresponding to the new compounds (ordered by Registry Number)
Microfilm the file
Update the permanent microfilm file of Registry Forms

REGISTRY SYSTEM CHART B

Flow Sheet for Compounds Selected from Biochemical, Industrial, and Physical Sections of Chemical Abstracts

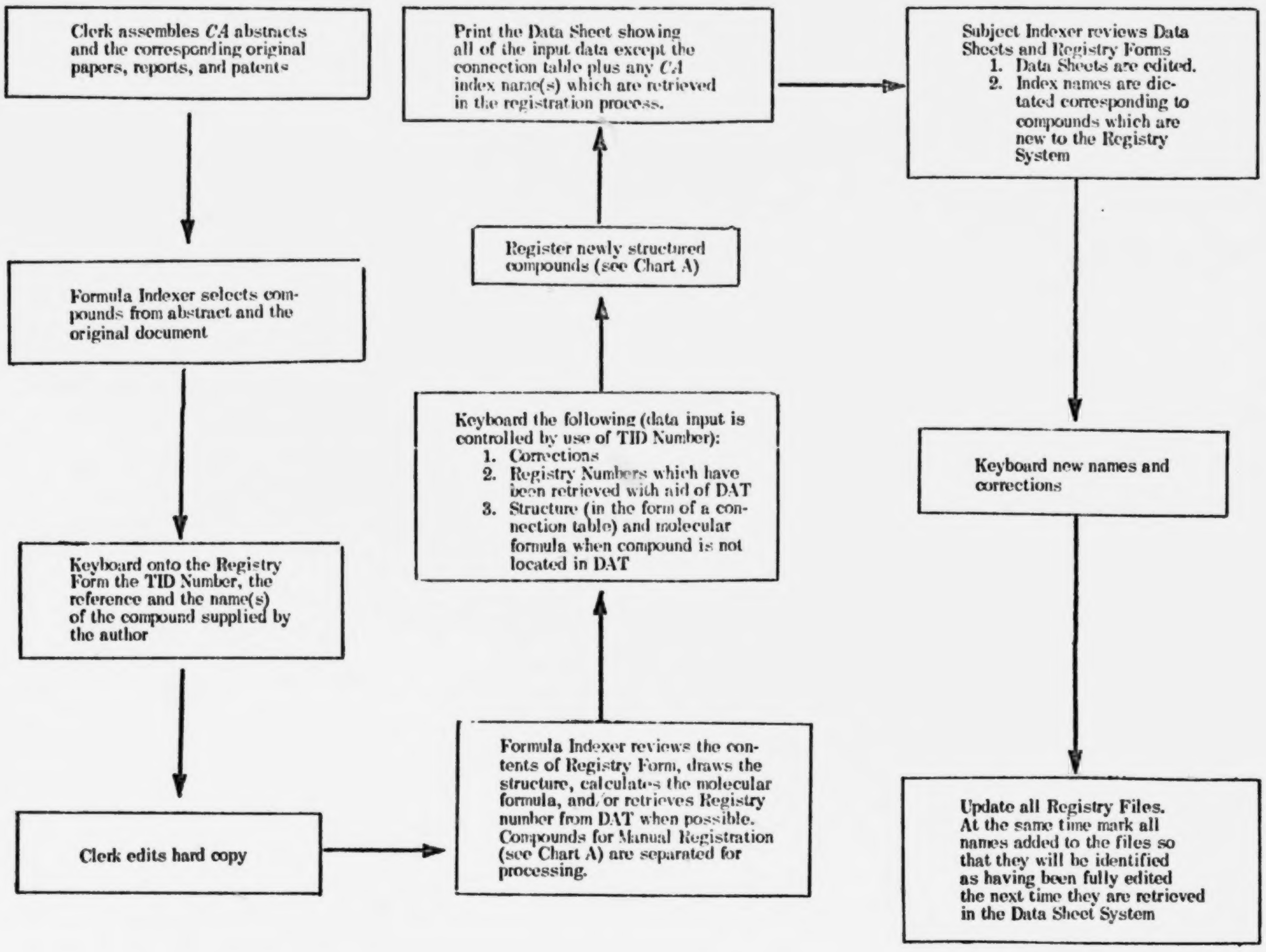
The registration operation is a batch process, with batches based on the organization of material in CA. The subject material indexed by the processes illustrated on this flow chart deals mainly with nonstructural data. The compounds described are most often well known and are associated with trivial names in the documents being indexed. Thus, after a compound has been selected for inclusion in an index entry, the author's name for the compound is automatically matched against the file to try to retrieve the CA index name(s) and the associated Registry Number for the compound. If the index name is retrieved, the corresponding connection table must already be on file. If the index name cannot be retrieved by machine or by use of a DAT, then the structure of the compound must be drawn and registered. Thus, the procedure illustrated here eliminates the need to redraw and rename the structures for all of those compounds which respond to the name-match technique.


Clerk assembles *CA* abstracts and the corresponding original papers, reports, and patents.
Subject Indexer identifies and dictates the subject index entries. The names for compounds assigned by the author in the document being indexed are used by the indexer. The abstract and very frequently the original document are used by the indexer in the indexing process.
Keyboard dictated entries
Computer-based match of names of compounds against the Registry Nomenclature File (See Chart A)
Print the Data Sheet showing all of the input data plus the *CA* index name(s), molecular formula, and Registry Number if obtained in the match of names.
Clerically proof Data Sheets.
Formula Indexer reviews the Data Sheets:
1. If the data checks, the sheets are passed on to the Subject Indexer.
2. Discrepancies are checked against DAT.
3. Registry Forms are prepared for compounds for which no name match (machine or DAT) is obtained.
4. Compounds for Manual Registration (see Chart A) are separated for processing
Keyboard the TID Number, the structure (in the form of a connection table), and the molecular formula.
Register newly structured compounds (see Chart A)
Subject Indexer reviews Data Sheets and Registry Forms:
1. Data Sheets are edited
2. Index names are dictated corresponding to compounds which are new to the Registry System
Keyboard new names and corrections
Update all Registry Files. At the same time mark all names added to the files so that they will be identified as having been fully edited the next time they are retrieved in the Data Sheet System

REGISTRY SYSTEM CHART C

Flow Sheet for Compounds Selected from Synthetic Sections of Chemical Abstracts

These are batch processes, with the batches based on the organization of material in *CA*. In this material names of the compounds are often not included in the documents being indexed. Thus, it is usually more efficient to process the material through the structure-drawing operation and then through the subject indexing operation. By this mode of operation, the CA index names are often retrieved and reviewed with no need for keyboarding of names.


Clerk assembles CA abstracts and the corresponding original papers, reports, and patents
Formula Indexer selects compounds from abstract and the original document
Keyboard onto the Registry Form the TID Number, the reference and the name(s) of the compound supplied by the author
Clerk edits hard copy
Formula Indexer reviews the contents of Registry Form, draws the structure, calculates the molecular formula, and/or retrieves Registry number from DAT when possible. Compounds for Manual Registration (see Chart A) are separated for processing.
Keyboard the following (data input is controlled by use of TID Number): 1. Corrections 2. Registry Numbers which have been retrieved with aid of DAT 3. Structure (in the form of a connection table) and molecular formula when compound is not located in DAT
Register newly structured compounds (see Chart A)
Print the Data Sheet showing all of the input data except the connection table plus any CA index name(s) which are retrieved in the registration process.
Subject Indexer reviews Data Sheets and Registry Forms 1. Data Sheets are edited. 2. Index names are dictated corresponding to compounds which are new to the Registry System
Keyboard new names and corrections
Update all Registry Files. At the same time mark all names added to the files so that they will be identified as having been fully edited the next time they are retrieved in the Data Sheet System

III. NUMBERS AND SOURCES OF REGISTERED SUBSTANCES

One of the main objectives of Task I was the buildup of useful files of chemical information through the registration of substances identified during the indexing process for CA, substances contained in several CAS internal reference files, and substances identified in specified published reference works such as the Colour Index, the Merck Index, and United States Adopted Names.

Registration of substances identified in indexing CA began with CA Volume 62 (January - June 1965), and continued with subsequent volumes through Volume 69 (July - December 1968), processing for which was in progress as of 16 March 1969, when contract NSF-C414 ended.*

Initial registration of the specified CAS Reference Files and published sources was undertaken during the first two years of Contract NSF-C414. Most of the reference files are routinely updated as pertinent new information is added to them. One of the files, the CAS Silicon File, was registered as a special project between July 1968 and January 1969. This file contains organic silicon compounds referenced in CA Indexes from Volume 1 through the present and in Beilsteins Handbuch der Organischen Chemie.

In addition to these Task I sources, the Registry System includes 40,375 compounds registered by CAS prior to the start of Contract NSF-C414, and 82,505 substances added to the files as a result of other projects.

Tables I through IV summarize the activity and status of the Registry as of 16 March 1969. Table I presents an overall summary of the size of the files. Table II gives the total number of registrations performed, while Table III shows the number of these registrations that have resulted in the addition of new compounds to the files. (This table also breaks down the number of compounds into machine and manual registrations.) Table IV shows the first source of names on the Nomenclature File. That is, for each source listed, the table shows how many new names that source contributed to the file.

---

* Registration will be completed under Contract NSF-C853.

TABLE I

REGISTRY STATUS SUMMARY REPORT

1.	SIZE OF FILES	
	a. Number of Different Compounds	1,074,319
	b. Number of Mixtures	5,232
	c. Number of Different Names on File	1,420,236
	d. Estimated Number of References	2,181,600
2.	NUMBER OF REGISTRATIONS (MACHINE AND MANUAL)	
	a. Substances New to File	1,079,551
	b. Substances Matching One on File	1,102,091
	TOTAL	2,181,642
3.	SOURCES OF REGISTRATION	
	a. CA Indexes	1,529,139
	b. Task I Files	215,180
	c. Task III	38,329
	d. Other	398,994
	TOTAL	2,181,642

TABLE II

Total Number of Registrations Performed through 16 March 1969

Source	Total Registrations	
A - Current Literature		
CA Volume 62	168,859	
CA Volume 63	203,895	
CA Volume 64	210,563	
CA Volume 65	218,041	
CA Volume 66	204,397	
CA Volume 67	210,575	
CA Volume 68	211,421	
CA Volume 69	101,388	
Subtotal		1,529,139
B - CAS Reference Files		
Alkaloid File	4,318	
Colour Index	15,019	
Drug File	1,622	
Fluorine File	59,814	
CA Formula Index Cross-References	3,260	
Lange *Handbook of Chemistry*	9,722	
Merck Index	21,266	
Pesticide Index	932	
CAS Reference File	439	
Ring Index (plus supplements)	30,204	
Silicon File	15,141	
SOCMA Handbook	6,768	
CA Specific Volume Cross-References	10,713	
Steroid File	1,540	
CAS Subject Index Cross-References	31,073	
Terpene File	1,829	
USAN (*United States Adopted Names*)	1,520	
Subtotal		215,180
C - Other Registration (Not Task I)		437,323
D - Total*		2,181,642

* Includes 40,375 compounds registered before 1 June 1965.

TABLE III

SUMMARY OF REGISTRATION

Source	Unique Substances Registered		
	By Machine	Manually	
A - Current Literature			
CA Volume 62	115,177	6,318	
CA Volume 63	131,958	3,262	
CA Volume 64	120,700	3,413	
CA Volume 65	106,553	7,634	
CA Volume 66	99,929	6,253	
CA Volume 67	99,530	4,162	
CA Volume 68	95,709	4,554	
CA Volume 69	54,798	234	
Subtotal			860,184
B - CAS Reference Files			
Alkaloid File	747	1,403	
Colour Index	3,270	4,190	
Drug File	335	2	
Fluorine File	42,255	3	
CA Formula Index Cross-References	952	20	
Lange Handbook of Chemistry	4,205	565	
Merck Index	6,027	2,456	
Pesticide Index	154	16	
CAS Reference File	175	29	
Ring Index (plus supplements)	23,473	-	
Silicon File	11,843	148	
SOCMA Handbook	6,480	140	
CA Specific Volume Cross-References	1,725	136	
Steroid File	348	311	
CAS Subject Index Cross-References	16,105	53	
Terpene File	517	372	
USAN (United States Adopted Names)	89	78	
Subtotal			128,622
C - Other Registration (Not Task I	82,505	8,240	90,745
D - Total*			1,079,551

* Includes all registration prior to 1 June 1965

TABLE IV

FIRST SOURCE OF NAMES ON NOMENCLATURE FILE

Source	Names on File 17 March 1969
A - Current Literature	
CA Volume 62	171,431
CA Volume 63	179,463
CA Volume 64	219,823
CA Volume 65	211,516
CA Volume 66	184,653
CA Volume 67	180,257
CA Volume 68	209,454*
CA Volume 69	110,975*
B - CAS Reference Files	
Alkaloid File	3,161
Colour Index	39,092
Drug File	2,559
Fluorine File	51,672
CA Formula Index Cross-References	3,409
Lange Handbook of Chemistry	11,898
Merck Index	29,980
Pesticide Index	3,388
CAS Reference File	3,293
Ring Index (plus supplements)	21,512
Silicon File	15,591
SOCMA Handbook	27,292
CA Specific Volume Cross-References	18,220
Steroid File	2,682
CAS Subject Index Cross-References	62,827
Terpene File	3,202
USAN (United States Adopted Names)	2,518
Other Registration (Not Task I)	132,631

* Processing incomplete as of 17 March 1968

IV. REGISTRY INPUT METHODS

The Chemical Registry System requires the routine translation into machine language of structural information, names, references, and control information. In this large-scale system, it is of great economic importance to develop efficient procedures for creating this data in machine-readable form while at the same time assuring the reliability of the information that enters the files.

During the early stages of Contract C414, CAS experimented with several methods of recording structural and nonstructural information in machine-readable form. These tests involved equipment evaluation, the establishment of keyboarding conventions, and the continued refinement of data flow paths in the system.

The following sections describe the various input methods that have been experimented with as well as some improved procedures which have been instituted to save time in the processing.

A. Input of Structural Information

1. Manual Generation of Connection Tables Followed by Keypunching.

The first input method used for the registration of structures in the system was the manual clerical generation of connection tables followed by the keypunching of connection-table data into cards. In this process, each nonhydrogen atom in the structural diagram is numbered by a clerk, and then a connection table is written which lists each atom by number, indicating the atom to which each is bonded and the types of connecting bonds. Finally, each rank of the table for input to the computer is keypunched.

This manual procedure was used during the first nine months (June 1965 through February 1966) of the contract.

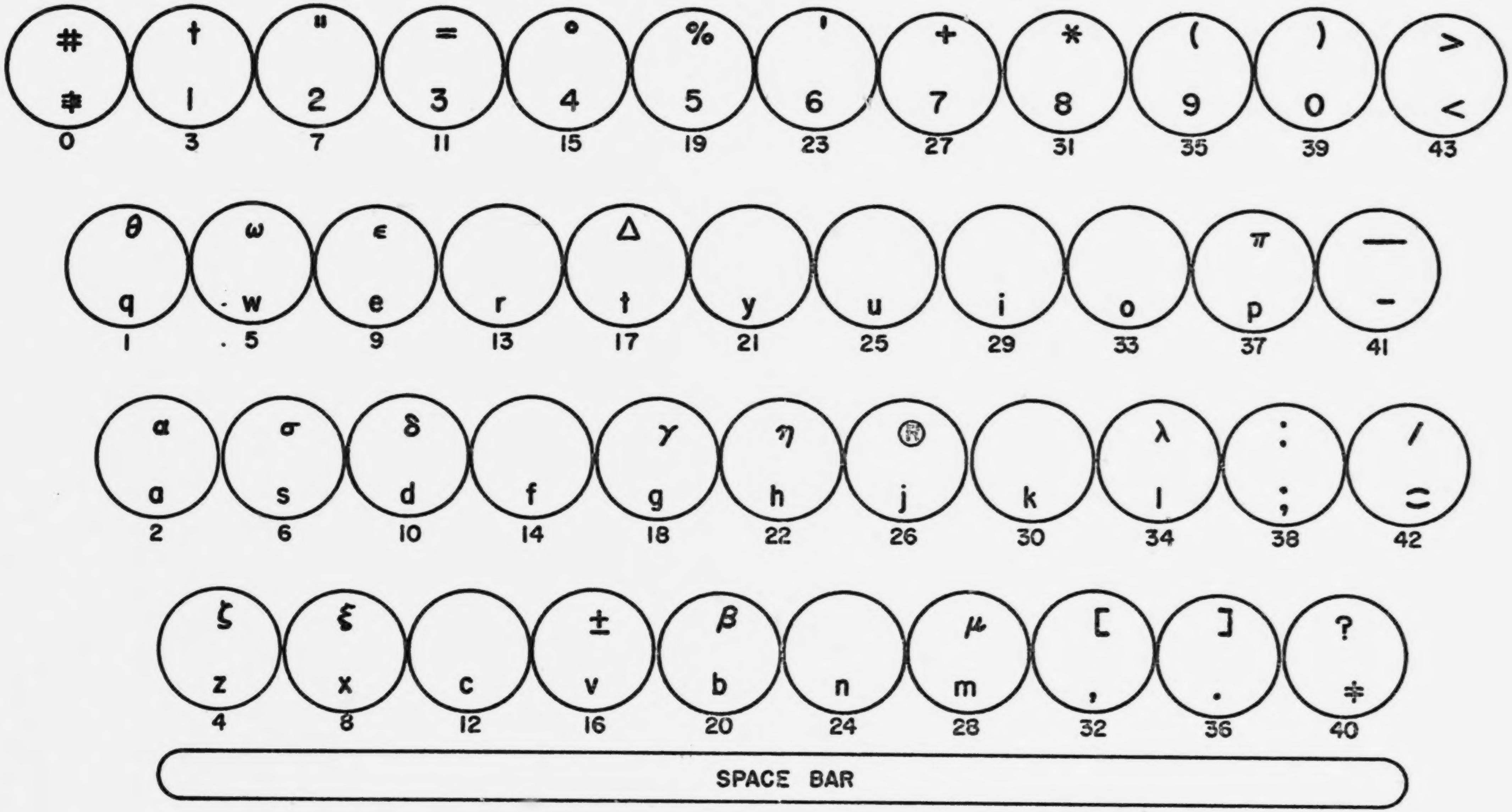
2. Direct Keypunching of Connection Tables.

Beginning in October 1965, CAS began testing a direct keypunching of connection tables. In this modification, the manual generation of connection tables is eliminated and they are keyboarded directly from the structure, which the clerk numbers before she keypunches the information.

This method has an obvious advantage over the first method in the elimination of a processing step. However, tests were conducted to determine whether this advantage was outweighted by decreased output and increased errors in the connection tables. The experiments showed that direct keyboarding of connection tables created a net decrease in the cost of input.

3. Direct Typing of Connection Tables

In September 1965, CAS began testing the Dura Mach 10 paper-tape-punching typewriter as a direct keyboarding device for connection tables. This typewriter produces both hard copy and a punched paper tape in which each typed character is coded in machine language. In using the Dura typewriter to generate connection tables, each line of data is typed on the worksheet appended to the bottom of the Registry form. To ensure the most efficient operation, CAS designed a special keyboard for this typewriter. Figure 1 illustrates this keyboard.



# ‡ 0
† 1 3
" 2 7
= 3 11
° 4 15
% 5 19
' 6 23
+ 7 27
* 8 31
(9 35
) 0 39
> < 43
θ q 1
ω w 5
ε e 9
r 13
Δ t 17
y 21
u 25
i 29
o 33
π p 37
— - 41
α a 2
σ s 6
δ d 10
f 14
γ g 18
η h 22
Ⓡ j 26
k 30
λ l 34
: ; 38
/ ⊃ 42
ζ z 4
ξ x 8
c 12
± v 16
β b 20
n 24
μ m 28
[, 32
] . 36
? ‡ 40
SPACE BAR

Figure 1: CAS Text-Typing Keyboard. Special characters are in the uppercase positions as shown and in both positions on keys 0 and 40-43. Where no uppercase symbol is shown, the conventional capital letter is available. Such capitals are used as special "flags" in most text-typing operations. Uppercase is indicated by the double dagger (‡) (key 40) preceding the lowercase letter.

The use of paper-tape-punching typewriters requires the conversion of data from punched paper tape to magnetic computer tape. Initially, the only equipment available to accomplish this conversion required two conversion steps: first, conversion from paper to punched IBM cards using a Dura Converter, then the transfer of data from cards to magnetic tape using the computer.

Rather than use this inefficient method, CAS rented an early production model of the Digi-Data paper-tape to magnetic-tape converter, which accomplished conversion in a single step.

After some initial difficulaties with the hardware, CAS found this converter to be satisfactory. Moreover, conversion was accomplished at about 37% of the cost of the previous method. However, since the average net cost of input by this method was found to be higher than that incurred with the Mohawk 1101 Data Recorder (see below), its use was suspended in February 1966.

4. Mohawk 1101 Data Recorder.

Since the beginning of March 1966, CAS has been using the Mohawk 1101 Magnetic Tape Keyed Data Recorder for input of connection tables. The Mohawk 1101 is a type of keypunch which records directly onto magnetic tape, eliminating the need for punched cards or paper tape. Although its operation is almost identical to the keypunch, the Mohawk is slightly faster for certain operations (e.g., skipping or duplicating) because the tape transport speed is faster than card transport speed. There is no method, except computer printout, for producing hard copy on the Mohawk 1101, but this has caused no problems in the direct generation of connection tables.

5. Improved Error-Correction Technique.

Early in 1966, CAS instituted an improved, computer-supported error-correction technique for connection tables. In order to avoid re-keyboarding the entire table when one was rejected for error, CAS developed a pending routine that "saved" the rejected table and printed it for the clerk. Upon identifying the error, the clerk then entered only the ranks of the table that were in error, not the entire table. The corrected information, merged with the "saved" table, then re-entered the input cycle.

6. The Structure Typewriter.

In April 1967, a system of direct structure typing was instituted. A Mohawk 1181 Data Recorder connected to a specially modified typewriter mechanism* permits a clerk to copy hand-drawn structures, producing hard copy and computer-usable magnetic tape simultaneously. A companion computer program then converts the typed structure to a connection table for processing. The advantages of this method lie in the elimination of all clerical conversion steps required for the generation of the connection table, in the reduction of the average number of keystrokes per structure, and in the reduction of errors because the clerk is merely copying rather than translating it into another format.

This method has been found to be highly economical in both cost and time factors. Currently approximately 90% of all structures are input by this system. The remaining structures are generally complex compounds which are not readily adaptable to typing on the typewriter. These structures are registered by means of connection tables, as described in point 4 above.

---

* See Mullen, J. M.,""Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures." J. Chem. Doc. 7, p. 88-93 (1967). Figure 2 shows the layout for this typewriter keyboard.

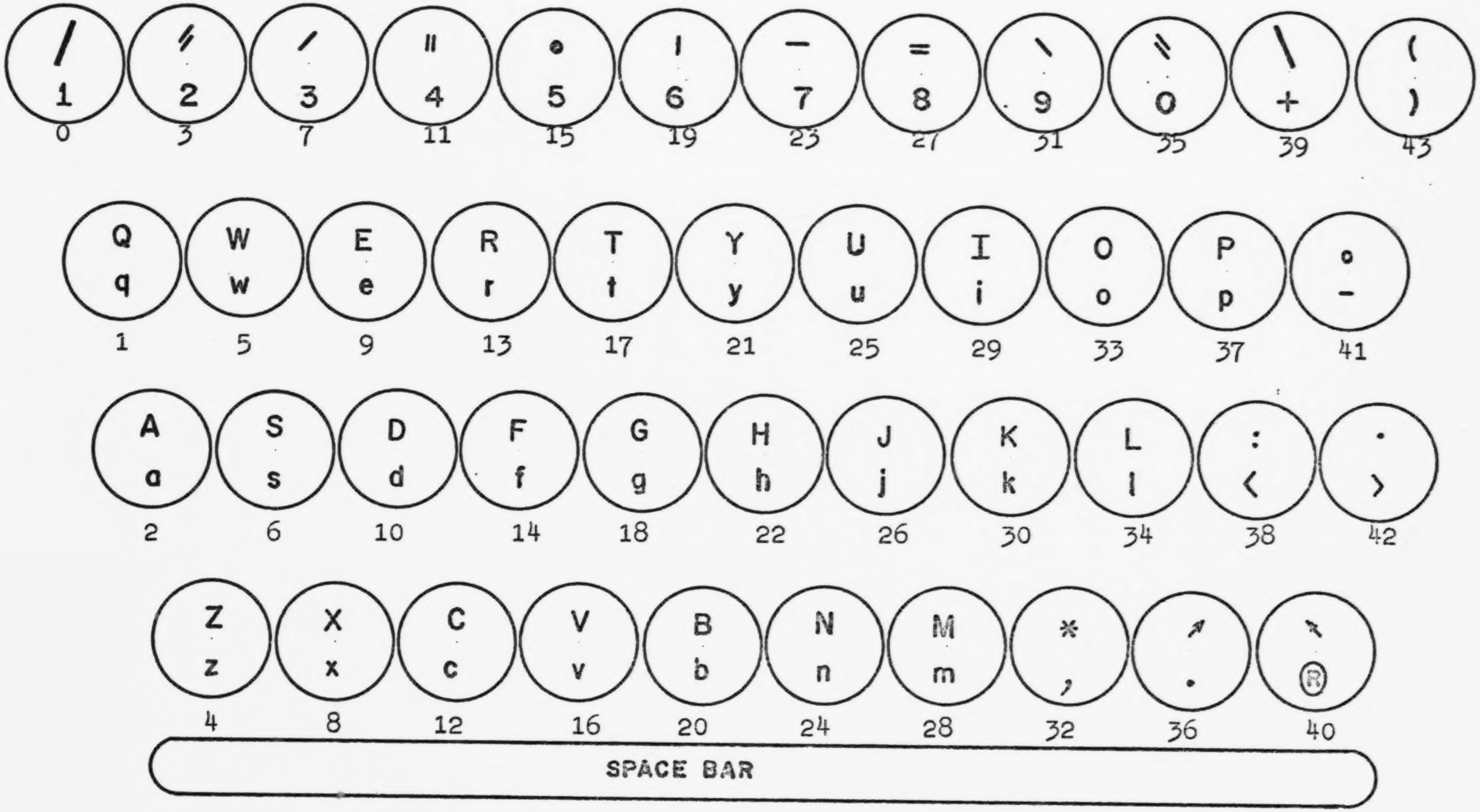

1 0
2 3
3 7
4 11
5 15
6 19
7 23
8 27
9 31
0 35
+ 39
( ) 43
Q q 1
W w 5
E e 9
R r 13
T t 17
Y y 21
U u 25
I i 29
O o 33
P p 37
° - 41
A a 2
S s 6
D d 10
F f 14
G g 18
H h 22
J j 26
K k 30
L l 34
: < 38
· > 42
Z z 4
X x 8
C c 12
V v 16
B b 20
N n 24
M m 28
* , 32
. 36
Ⓡ 40
SPACE BAR

Figure 2: CAS Structure-Typing Keyboard. Special bond symbols are shown on the top line of numerals. In addition to these special symbols, the typewriter is also equipped with keyboard-controlled platen movement both forward and backwards, and with half-line spacing.

7. The Use of Input Shortcuts for Structural Information.

The routine processing of structural diagrams for registration involves the repeated handling of many structure fragments that are common to several compounds. CAS has developed input shortcuts to reduce the processing time and the chance of error in handling such fragments by a system whereby a group of atoms can be handled as if it were a single atom. One such symbol -- Ph for phenyl -- was used at the start of the contract. Since then, additional shortcuts have been developed. The computer program recognizes each symbol and "expands" the record to the full set of atoms and bonds represented. Thus, the same computer record results whether or not the symbol is used.

B. Input of Nonstructural Information

There are two basic computer-based systems for entering nonstructural data into the computer files -- the Name Matching System and the Data Sheet System.

1. Name Matching Capability.

With the large numbers of registered compounds on file, it is sometimes more efficient to update the files by retrieving the Registry Number through a compound's name rather than its structure. The name of the previously registered compound can be matched against the names already on file, the Registry Number retrieved, and the new bibliographic information on the compound can be added to the file via the retrieved Registry Number.

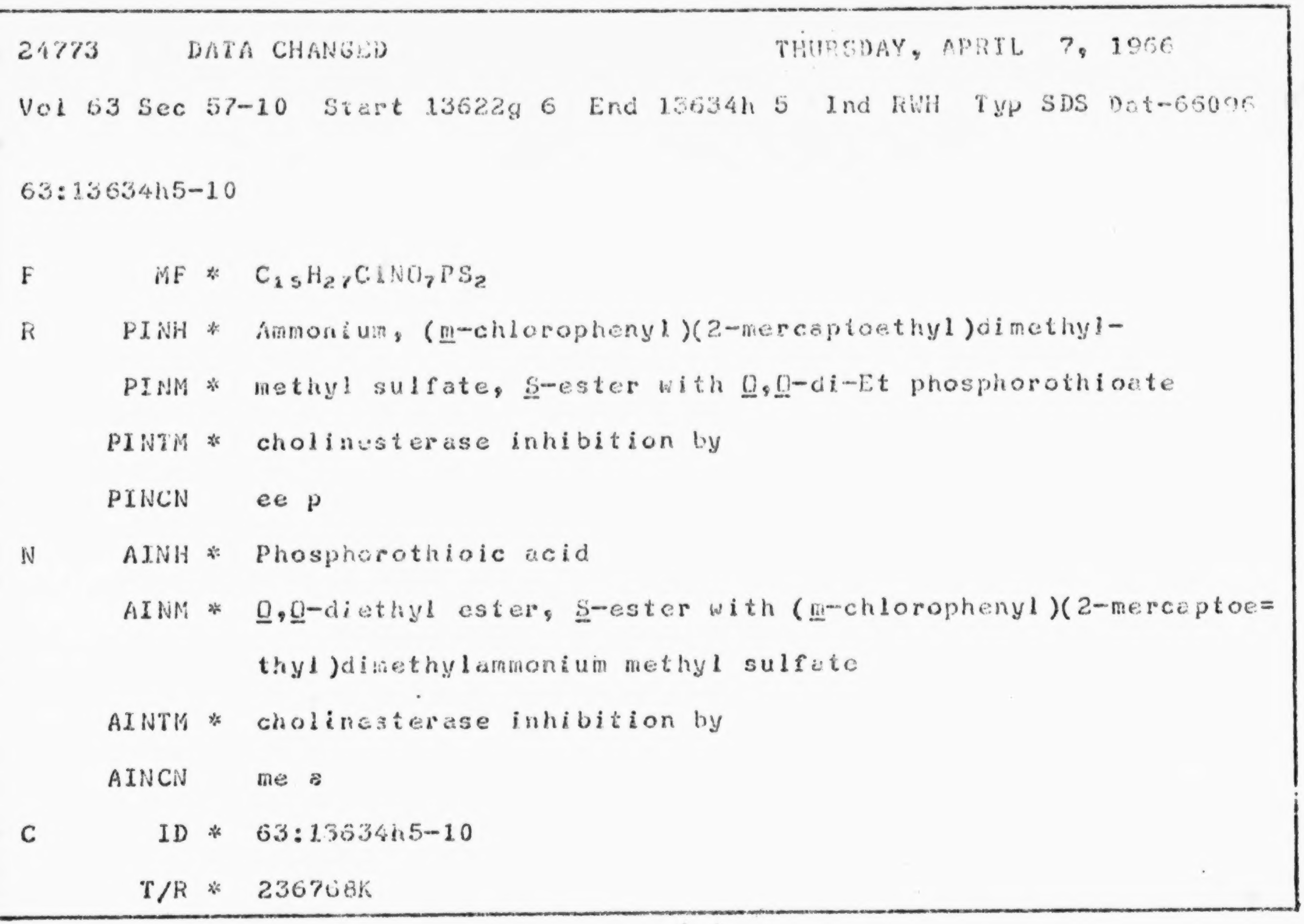
Manual name matching was used during the first nine months of the contract by chemists employing the Desktop Analysis Tools. However, this method became inefficient as the files grew larger. Therefore, during the first quarter of 1966, CAS developed and began using a computer-based name-matching capability by which names of compounds being processed for registration may be compared with already registered compounds. The chemist must still be involved in the process to recognize ambiguous names. When an exact match is achieved, the Registry Number, the molecular formula and the _CA_ index name (if available) can be retrieved from the Nomenclature and Bibliography Files.

In routine operations, the computer-based Name Match System is used for those segments of the input which experience has shown to contain a relatively large percentage of compounds with common trivial names. These segments include the Biochemical, Industrial, and Physical Sections of _CA_ (see Registry System Flow Chart B, page 8).

2. The Data Sheet System

Beginning in late 1965, CAS began developing the Data Sheet System, so called from the special work sheets which are printed by the computer to facilitate chemists' review of data being input to the files. (See Figure 3.) The Data Sheet System is the fundamental tool whereby nonstructural information -- primarily names and references -- is developed in machine language, edited, corrected where necessary, manipulated by the computer, and finally output for updating the Registry Files and for use in _CA_ indexes.

The first requirement for this system was the installation of equipment with which clerks could, in a single keyboarding, produce data in both hard-copy and machine-language forms. The standard electric typewriters that had been used to generate _CA_ index cards were replaced first with



24773 DATA CHANGED THURSDAY, APRIL 7, 1966

Vol 63 Sec 57-10 Start 13622g 6 End 13634h 5 Ind RWH Typ SDS Dat-66096

63:13634h5-10

F MF * $C_{15}H_{27}ClNO_7PS_2$

R PINH * Ammonium, (m-chlorophenyl)(2-mercaptoethyl)dimethyl-

PINM * methyl sulfate, S-ester with O,O-di-Et phosphorothioate

PINTM * cholinesterase inhibition by

PINCN ee p

N AINH * Phosphorothioic acid

AINM * O,O-diethyl ester, S-ester with (m-chlorophenyl)(2-mercaptoe=

thyl)dimethylammonium methyl sulfate

AINTM * cholinesterase inhibition by

AINCN me s

C ID * 63:13634h5-10

T/R * 236768K

Figure 3: Example of a Computer-Produced Data Sheet. The top line of information shows the sequential number of the Data Sheet, the type of worksheet ("Data Changed in the example; other possible types include "New Worksheet," and "Entry Killed"), and the date the Data Sheet was produced. The second line of information gives the CA identification, including the volume and section numbers and the start and end of the batch of abstracts being processed; the indexer's initials; the typist's initials, and the Julian-form date of original typing. The third line gives the CA reference for the data that follow.

In the following data, the left-hand column contains single-letter codes that indicate the beginning of each new field of data: the F field for the molecular formula (MF), the R field for the Preferred CA Index Name (PIN), the N field for the Added CA Index Name (AIN), and the C field for control information. Each of the name fields may contain the name heading (H); the name modification (M); Stereochemical identification (S) for the name (not illustrated); the text modification (TM) used in the CA Subject Index; and the corner note (CN), a coded production aid used in index generation. The Control information field includes the CA identification (ID), and the TID number or Registry Number (T/R) of the compound.

paper-tape-punching electric typewriters, then with Mohawk Data Recorders. This change eliminated a separate keypunching operation that had previously been used to generate names in machine language, and it allowed the CA index names to be captured on tape as a by-product of the index transcription process.

Beginning in January 1966, CAS began large-scale operation of the Data Sheet System by installing new computer programs and changing the data flow in the Compound Registry System to provide computer support for the review of nonstructural information by chemists. (See Appendix E.) With this extension, the Data Sheet System ensures that no information enters the master Registry File until it has been reviewed by an appropriate staff member and corrected where necessary. At the same time, the system prevents redundant editing of names and references entering the computer file.

The role of the Data Sheet System in Registry operations is illustrated in Registry System Flow Charts A, B, and C. In outline, the Data Sheet System in its present form operates as follows: Information is typed on tape-generating typewriters, thus simultaneously creating a hard copy for review and a magnetic tape for input to the computer. The information is processed by the computer into a work-pending file, and on call from a CAS chemist, the data are printed on computer-generated data sheets. These are proofread against the original hard copy as a check on reliability, and the data are then reviewed for chemical accuracy. When the necessary corrections have been made, keyboarded, and entered into the work-pending file, the chemist approves the transfer of the reviewed data to the master Registry Files. At the time of transfer, the computer "flags" the data as having been edited; this flag serves the purpose of preventing unnecessary re-review of the data at some later time.

The Data Sheet System also makes it possible to eliminate much repetitive keyboarding because the system provides the typist with special codes that signal the computer to repeat specifically identified information which need not be retyped. Like the input shortcuts for structural information, these short codes are expanded by the computer into their full representation. Thus, failure to use the code will not affect the appearance of data in the computer file. A brief example of such "dittoing" is given below for the Id code, which reduces keyboarding for lists of CA inverted systematic names. The Id code signals the repetition of that portion of a CA index name which precedes a comma followed by a space, e.g., the comma of inversion. This dittoing feature reduces keystrokes in the typing of alphabetized lists of names, where one parent may have several substituents grouped at one point on the list. Other dittoing codes enable the typist to signal the repetition of data from one sub-field in a data sheet to another sub-field in the same sheet or from one sub-field in a data sheet to the corresponding sub-field in the following sheet(s).

Name	Typed as
Galactopyranose, 2-acetamido-2-deoxy-	Galactopyranose, 2-acetamido-2-deoxy
Galactopyranose, 3-amino-1,6-anhydro-3-deoxy	Id3-amino-1,6-anhydro-3-deoxy
Galactopyranose, 1,6-anhydro-	Id1,6-anhydro-

V. IMPROVEMENTS AND EXTENSIONS IN THE REGISTRY

Since the basic registration capability was established in 1965, CAS has continually worked to extend the scope and efficiency of the system. This section describes many of these improvements made after initial operations of the Registry had begun in 1965. The extensions of the system undertaken during the redesign and reprogramming of the System for the IBM 360 computers are described in Section VII. of this report.

A. Extension of Registry Algorithm

Very early in the Contract work, the Registry Algorithm was modified to detect patterns of symmetry in a structure and to utilize such symmetry to prevent the unnecessary generation of equivalent connection tables as "candidates" for the unique connection table. (Such candidate tables are generated and compared with each other to permit the hierarchical sorting and selection of the one unique table). As a result of this modification, the average time required for the computer generation of unique connection tables was reduced about 20%.

B. Computer-Checked Temporary Identification Numbers

Temporary Identification (TID) numbers are used in the Registry System to identify each structure at input until a Registry Number is assigned to the substance or retrieved for it. With the large-scale input required under Task I, CAS adopted computer-checkable TID numbers in June 1965. These TID numbers which are preprinted on Registry Sheets, consist of six digits and a letter suffix. The digital portion is evenly divisible by seven, thus providing machine checkability. At the time this TID number format was adopted, computer routines to check the TID numbers were also instituted.

C. "Dot-Disconnected" Convention

The "dot-disconnected" convention for structures and molecular formulas of salts, clathrates, addition compounds, etc. was adopted during the first six months of Task I. With this convention, a compound is represented as the structure of the "parent" portion, followed by a dot and the structure of the salt forming portion(s). Each moiety following the first (there may be several moieties, each separated by a dot) carries a numerical prefix to indicate its proportion to the first (parent) moiety. This coefficient may be an "x" if the ratio is unknown. The molecular formulas of "dot disconnected" compounds are treated analagously to the structural diagrams.

This treatment saves storage space in the registry, allows the grouping and automatic cross referencing of related compounds (for example, salts with the same parent) and permits registration of those partially-determined structures in which the ratio between two fully defined moieties is unknown.

D. Extension of the Registry to New Classes of Compounds

At the start of Task I, CAS registered only organic (i.e., carbon-containing) compounds for which fully defined structural diagrams could be drawn (excluding coordination compounds). Since then structuring conventions and computer programming have been accomplished to extend registration to several additional types or classes of compounds. In general, the

CAS policy has been to accumulate nonregisterable compounds for study, and when enough examples of any one type are available, to establish structuring conventions and do the necessary computer programming to permit registration of these additional classes. Under Contract C414, the following types of compounds were added to those that could be machine registered:

1. Machine Registration of Free Radicals

In October 1965, CAS decided to register free radicals (that is, compounds in which one or more atoms bear an unpaired electron) by specifying in the connection table the abnormal valence of the atom(s) bearing the unpaired electron(s). Specifying the valence results in an override of the computer editing technique for the valence of atoms, and permits the compound to be registered with the specified abnormal valence.

2. Compounds Labeled Unspecifically with Isotopes

In several areas of chemistry, it is common to "label" compounds with isotopes in order to permit tracing the compound in an experiment. The Registry System from the beginning has been able to handle isotopically labeled compounds when the structural positions of the labeling isotopes are known. In September 1965, an extension was made in the system to permit registration of labeled compounds when the structural positions of the isotopes are unknown or when the number of atoms of the labeling isotope is unspecified.

This extension involves the use of textual descriptors for unspecifically labeled compounds; each descriptor consists of the elemental symbol and mass number of the isotope, together with an arabic numeral or an "X" indicating the number of atoms of the isotope. For example, a compound labeled with two atoms of carbon 14 would be registered with a textual descriptor 2C-14. If the number of labeling atoms were unknown, the descriptor would be XC-14.

3. Oligomers

During the second quarter of 1966, CAS developed the structuring conventions to permit the registration of oligomers. These compounds are polymers in which a fully defined unit is repeated a known number of times, but for which the interconnection between units is usually unknown. An example is the propylene pentamer $(C_3H_6)_5$. Oligomers are registered on the basis of their structural units, with a text descriptor (dimer, trimer, etc.) indicating the number of repetitions of the unit.

4. Inorganic Compounds

Inorganic compounds are defined as those that do not contain carbon. As a result of study of this class of compound during the early registry operations, CAS staff determined that inorganic compounds could best be registered by adopting the bonding conventions used for organic compounds, thus retaining full structural specificity. The accumulated file of inorganic compounds (i.e., the inorganic compounds from CA Volume 62 on and from the CAS reference files) were registered between June 1967 and February 1968. Since then, inorganic compounds have been routinely registered.

5. Coordination Compounds

For the Registry, coordination compounds are defined as compounds in which the number of attachments to a central atom (usually a metal atom, but not necessarily) exceeds the generally accepted valence of the central atom. In general, structures are represented by direct connection (bonds) between the central atom and the attached groups (ligands). The oxidation state of the central atom is represented by a charge which may be positive, negative, or zero. A ligand may also be neutral, positive, or negative; its charge is indicated on the atom directly attached to the central atom. The coordination number, represented by the number of attachments to the central atom, and its oxidation state are significant data for coordination compounds of that element, and thus must be included in the recorded data.

In preparation for the Eighth Collective Indexing Period (1967-1971) CAS staff defined new structuring and naming conventions for coordination compounds, permitting their consistent treatment in indexing and registration. Coordination compounds having six or fewer attachments to the central atom were input to the Registry System as it operated for the 7010 computers. The reprogramming of the Registry Structure System for the IBM 360 allowed input of structures having as many as 15 attachments to any one atom. In addition, the number of charges and the number of attachments to a given central atom are automatically edited in the 360 system.

6. Chemical Elements

The chemical elements themselves were not initially processed because the Registry System was programmed not to accept any "graph" that involved only a single nonhydrogen atom. This provision, designed as an input editing check, proved in practice to be of little value. Therefore, a program adjustment has been made to permit the registration of the elements, their isotopes, and allotropes. The elements and their isotopes and allotropes from CA Volumes 62-66 were registered between June 1967 and February 1968; the less common isotopes and allotropes continue to be processed for current volumes of CA.

7. Boranes and Carboranes of Known Structure

Boranes and Carboranes, which contain molecular skeletons of interlinked boron atoms or interlinked boron and carbon atoms, respectively, exhibit unusual aspects of structural representation. Based on a study of such compounds, and aided by nomenclature conventions defined by the ACS Council Committee on Nomenclature, CAS staff developed appropriate structuring conventions and input procedures to permit these compounds to be routinely registered. These procedures were available for use in February 1968.

8. Partially Determined Structures

Partially determined structures are substances about which some structural information is known, but for which full structural definition is not available. Two types of partially determined structures -- "dot disconnected" fragments with unknown ratios, and oligomers -- are now registered, as described above. In addition, CAS has defined the structuring conventions needed to register several additional classes. These include compounds in which two or more fully defined fragments are connected in an unspecified way. Some limitation may be given about the connection -- for instance a substituent may be known to be attached to a ring in a ring-containing compound but not to any of the chain portions of the compound. Another type of partially determined structure includes one fully defined structural portion and one or more attached fragments represented only as summation molecular formulas.

The procedures developed for handling these structures allow the computer record to reflect as much specificity as is present in the original document. Implementation of these procedures awaits completion of computer programming.

9. Polymers

Polymers present problems in registration because these substances are made up of recurring structural units, each of which may be regarded as derived from a specific compound, or monomer. The number of units is usually large and variable, a given polymer sample being characteristically a mixture of structures with different molecular weights. Since the beginning of Contract C414, CAS has been studying the ways polymers are reported in the literature in order to define efficient and meaningful registration methods. These studies show (see Appendix G) that polymers can be registered:

- a. On the basis of their structural repeating units (SRU's) with monomer information if available;
- b. In terms of their monomers, if no SRU information is given;
- c. By nonstructural data (name, application(s), generic types), if no structural information is given.

Procedures for registering polymers have been developed; implementation of these procedures awaits the completion of computer programming.

E. Improved Text Descriptor Processing

Those aspects of a chemical structure not incorporated into the two-dimensional structural diagram -- for example stereochemistry -- are handled in the Registry System by means of text descriptors. Each such descriptor is a string of letters, numbers, and/or punctuation that is an integral part of the structural record. Two compounds that possess the same two-dimensional structure but differ in text descriptor will be assigned different Registry Numbers. To assure accuracy and consistency in this process, all such sets of compounds with like two-dimensional structures but different descriptors are reviewed in context by a chemist before final acceptance by the computer.

1. Standardized Descriptors for Alkaloids, Carbohydrates, Steroids, and Terpenes.

As the Registry System has matured, CAS has developed standardized and somewhat codified stereochemical descriptor conventions for four groups of compounds: alkaloids, carbohydrates, steroids, and terpenes. In general such descriptors are made up of abbreviations of the CA Preferred Names for the parent compounds, plus alphanumeric prefixes indicating stereochemistry of specified nodes of the structure. Each name implies the stereochemistry at certain nodes in the structure; thus, the prefixes need specify stereochemistry only for the nodes not implicit in the name. As an example, consider the alkaloid compound with the CA index name Crinan-3α-ol,1β,2β-epoxy-7-methoxy. "Crinan" is the parent compound and nodes 1, 2, and 3 have stereochemistry that must be specified. The descriptor for this compound is 1B,2B,3A-CRINAN.

2. Automatic Editing of Text Descriptors

In order to reduce the amount of professional attention required for the resolution of text descriptors, a table of acceptable descriptors has been established for computer checking (see Appendix F). Descriptors that exactly match one on the list will be accepted by the computer without review by a chemist. Other descriptors can be entered in the file only when a chemist "flags" the descriptor to indicate that it is acceptable.

VI. DESKTOP ANALYSIS TOOLS

Contract C-414 calls for the development of Desktop Analysis Tools (DAT's) which, by functioning as specialized indexes to material in the computer files of the Chemical Registry System, eliminate some of the costly reprocessing of data that would be required if structures were completely reprocessed and re-registered each time they were encountered. Under Task I, CAS has developed and continually improved Desktop Analysis Tools that are computer-produced compilations of chemical names that have been entered into the system. These tools are used by clerical and chemical staffs to link the name of a compound with its Registry Number, and are thus useful in manual name-matching.

The first Desktop Analysis Tool produced was a single volume containing names ordered in strict alphanumerical sequence beginning with the first character in a name. While this ordering was straightforward and consistent, it resulted in the separation of some names that chemists were accustomed to seeing together. For example, trans-stilbene was alphabetized with the t's, whereas cis-stilbene was alphabetized with the c's. Moreover, as the Registry Files grew larger, the DAT's did also, making them more cumbersone to use and requiring some method for updating. Therefore, during the first year of the contract, CAS developed methods for publishing the DAT's in different volumes based on the type of name, for creating periodic supplements to the DAT's (each supplement containing only material added to the files since the last full DAT was issued), and for arranging names more nearly in the order expected by chemists.

The requirements of Task III of Contract C414, together with general work on improved input and correction procedures for names, resulted in a greatly improved capability for producing DAT's. The Task III DAT's are described in the "Final Report to the National Science Foundation on Contract NSF-C414, Task III," and in the special CAS report to the NSF: "Policies and Procedures Governing the Compiling of the Desktop Analysis Tools for the Common Data Base."

Some of the generalized procedures for the improved handling of chemical nomenclature, which affect not only the DAT's but also the entire file of nomenclature, are described below. These computer-based checking procedures support the keyboarding and input procedures for chemical names, helping to assure consistency in the file and reducing the need for clerical and chemical review.

A. Italicization

The editing program developed at CAS automatically italicizes any single Roman letter surrounded by punctuation characters. It also italicizes a single alphabetic character that begins or ends a name, excluding small capital letters. The lower-case letters "d" and "t," used to cite the hydrogen isotopes deuterium and tritium, are also recognized and italicized, as is the capital letter "H" for hydrogen under certain conditions. Also italicized are the alphabetic characters used within ring system names when they occur within the ring fusion brackets, unless they are immediately preceded by numerals.

In addition to the above, there is a list of 54 character strings that are automatically italicized when surrounded by punctuation. This list includes such stereochemical descriptors as "cis-," "trans-," and "erythro-." The same list is used to designate terms which are not to be capitalized when they occur at the beginning of a compound name and are followed by punctuation. This list is open-ended, and additional descriptors can be added as they are incorporated in chemical nomenclature.

B. Capitalization

Analyses of computer nomenclature files indicate that there are an average of 1.6 capital letters per name. The majority of these are the first letter of the basic name. The program automatically supplies this capital if it occurs in a sequence of two or more Roman letters. It will not capitalize the letter if it is within any of the character strings to be italicized.

Other capitalization is also automatically supplied by the program. The single alphabetic character which occurs most frequently as a capital within a name is the letter "H," usually to indicate hydrogen atoms in ring systems. The computer program identifies this letter in context as a capital "H" followed by any punctuation character and preceded by (1) the punctuation character, "prime," (2) by a numeral, or (3) by one of the letters a, b, c, or d, that, in turn, is preceded by a prime or a numeric character. The Roman numerals 1 through 10 are also identified and capitalized as are the stereochemical descriptors "R" and "S" cited within parentheses.

C. Checking of Punctuation Consistency

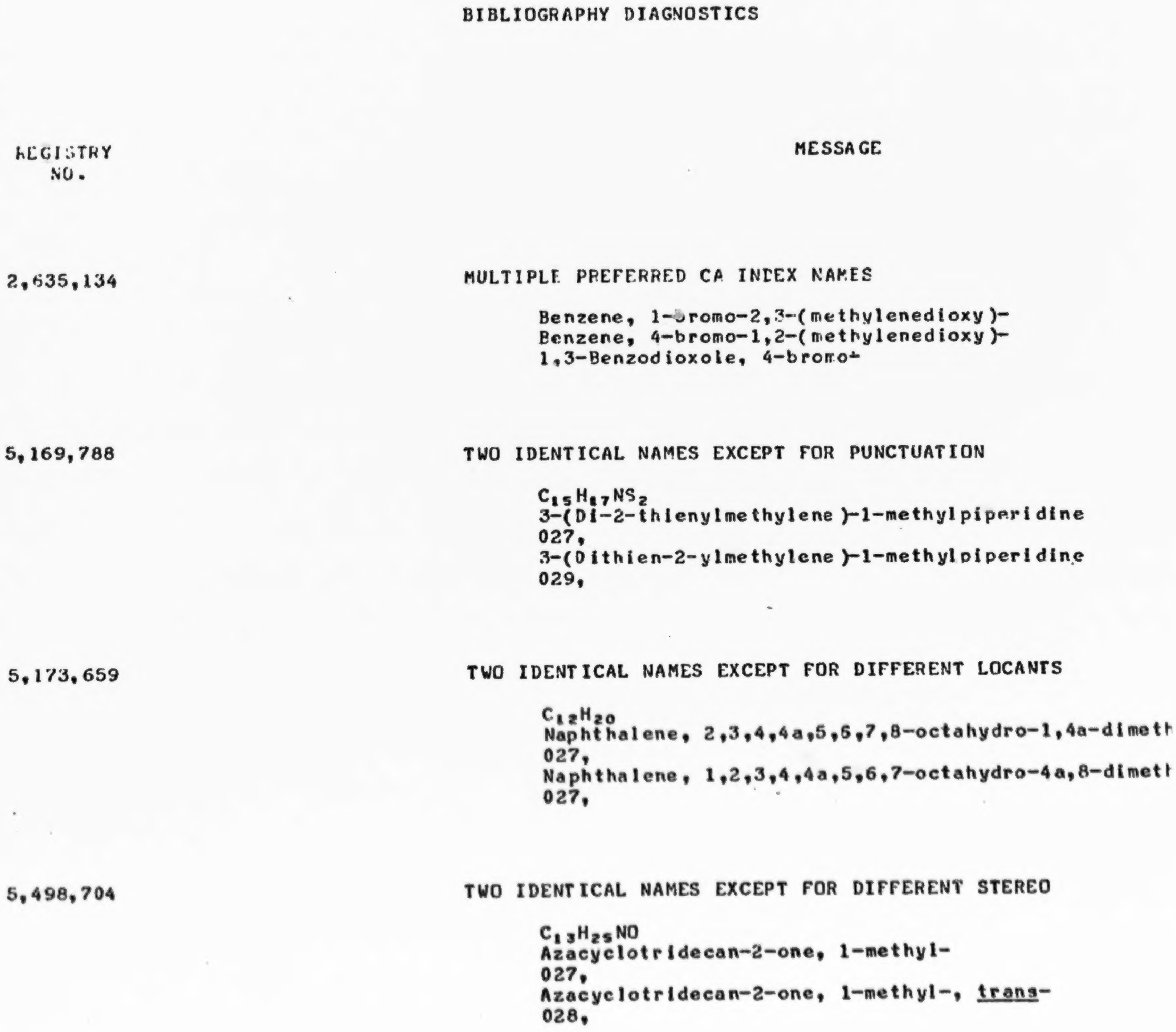
To prevent clerical errors from being recorded in the files, the computer is programmed to check for and correct (1) punctuation such as periods and commas incorrectly typed at the end of names, (2) for two blank spaces where only one is required, (3) for hyphens omitted between parentheses and brackets, and locants which immediately precede or follow such punctuation marks.

D. Elimination of Invalid Characters

Certain symbols such as the equality sign and quotation marks are used in Registry System operations as keyboarding conventions and do not imply their normal meanings. In the past, they have caused some confusion at printout. The computer program now checks for these characters and removes them when necessary.

E. Printing of Diagnostic Comments

Errors such as spelling, letter transposition, and significant punctuation cannot yet be automatically corrected. However, to reduce the number of names that must be manually reviewed, programs are in effect that detect the potential problem and display the questionable name along with a diagnostic comment (see Figure 4). This technique directs the professional's attention to the potential problem immediately, thereby reducing the time required to make editorial decisions. In addition, the computer files can also be updated without rekeyboarding the compound names.



```
                                BIBLIOGRAPHY DIAGNOSTICS


REGISTRY                                              MESSAGE
  NO.


2,635,134                     MULTIPLE PREFERRED CA INDEX NAMES

                                   Benzene, 1-bromo-2,3-(methylenedioxy)-
                                   Benzene, 4-bromo-1,2-(methylenedioxy)-
                                   1,3-Benzodioxole, 4-bromo-


5,169,788                     TWO IDENTICAL NAMES EXCEPT FOR PUNCTUATION

                                   C15H17NS2
                                   3-(Di-2-thienylmethylene)-1-methylpiperidine
                                   027,
                                   3-(Dithien-2-ylmethylene)-1-methylpiperidine
                                   029,


5,173,659                     TWO IDENTICAL NAMES EXCEPT FOR DIFFERENT LOCANTS

                                   C12H20
                                   Naphthalene, 2,3,4,4a,5,6,7,8-octahydro-1,4a-dimeth
                                   027,
                                   Naphthalene, 1,2,3,4,4a,5,6,7-octahydro-4a,8-dimet
                                   027,


5,498,704                     TWO IDENTICAL NAMES EXCEPT FOR DIFFERENT STEREO

                                   C13H25NO
                                   Azacyclotridecan-2-one, 1-methyl-
                                   027,
                                   Azacyclotridecan-2-one, 1-methyl-, trans-
                                   028,
```

Figure 4: Examples of Computer Produced Diagnostic Comments

F. Nomenclature Sort Key Program

In order to provide a listing of names ordered the way chemists expect to see them, CAS has developed a nomenclature sort key technique by which to arrange names.

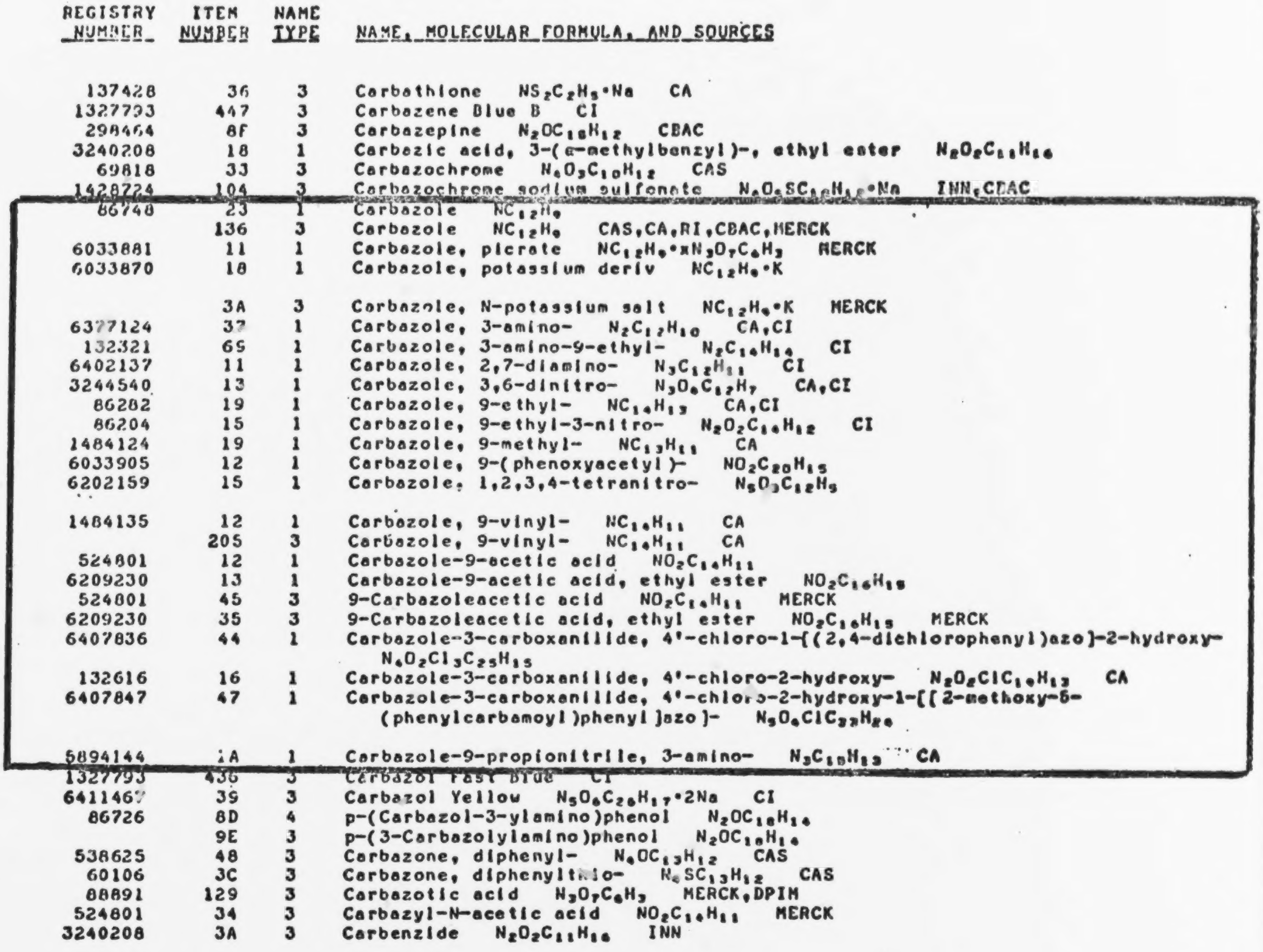
The method by which the Nomenclature Sort Key Program operates is as follows: Each name is divided (by internal computer routine) into four major portions: the parent, the substituents, the modifications, and the stereochemistry. Not every name will have all four portions, but, except for laboratory names, all will have at least an alphabetic parent. They may or may not possess the other portions described below. In a chemical name containing all possible portions, each field will contain an alphabetic string of characters and locants comprised of numerals or letters or both. The computer program is written such that each major field is investigated individually and in the order cited above. The Sort Key is generated so that, within each major field, the characters are rearranged with the letters first, followed by the locant. When the rearrangement is complete, the parts of the name are ordered in the following manner:

1. Alphabetic portion of parent
2. Locant portion of parent
3. Alphabetic portion of substituents
4. Locant portion of substituents
5. Alphabetic portion of modifications
6. Locant portion of modifications
7. Alphabetic portion stereo
8. Locant portion of stereo.

As an example of this rearrangement, the compound 1,2-Cyclohexanedicarboxylic acid, 3-hydroxy-, 1-ethyl ester, (+), would be rearranged as: cyclohexanecarboxylicacid 1 2 hydroxy 3 ethylester 1 (+). This rearranged name is the Sort Key. When a printout is called for, a program automatically alphabetizes the names by ordering on the Sort Keys. However, the printout contains not the Sort Key, but the properly edited name.

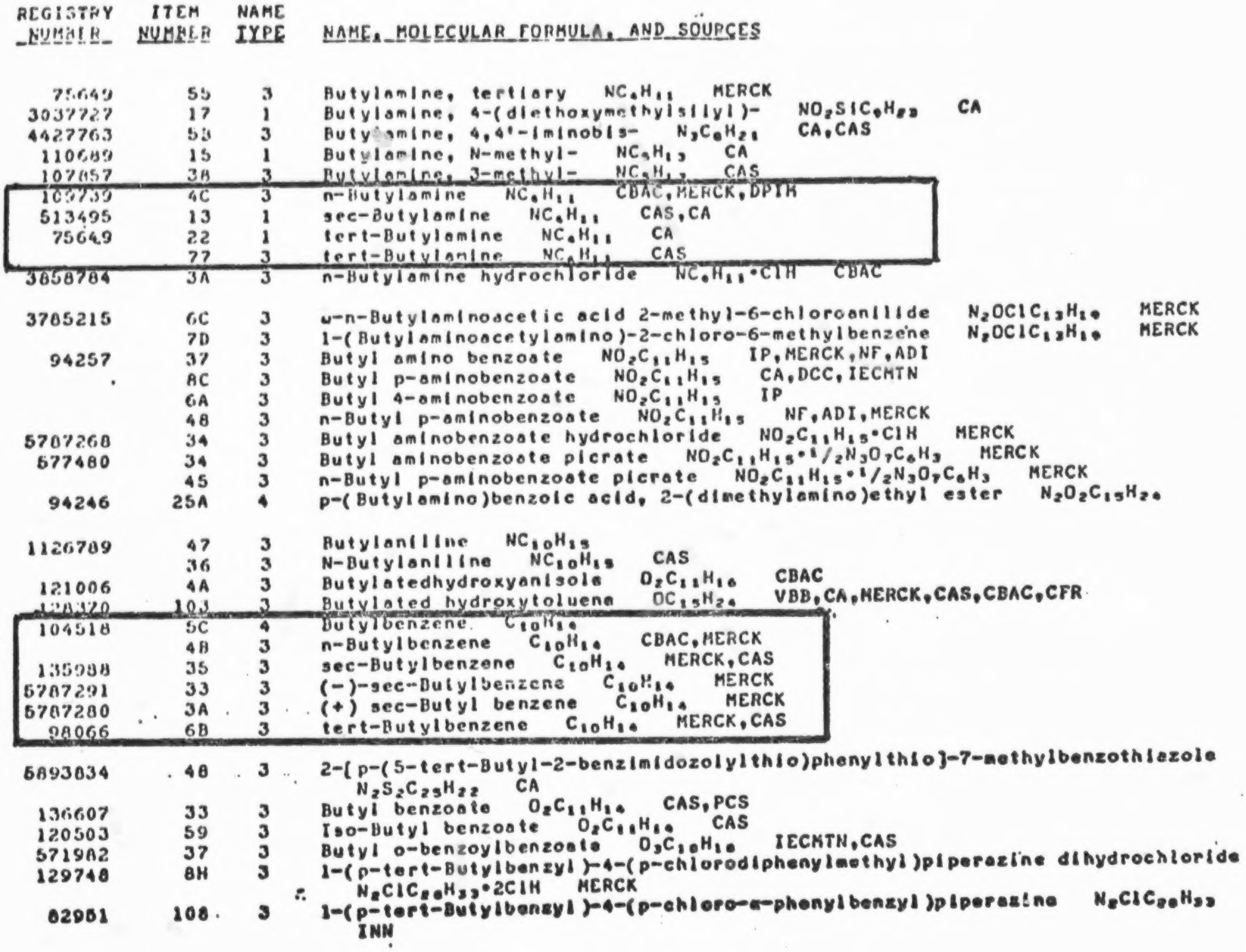
The above program accommodates virtually all types of chemical nomenclature, with one major exclusion -- laboratory names, which are usually composed solely of numerics and alphabetics. A second program is designed to sort this type of name. For example, the laboratory name B9963DEX would be sorted first by its numeric portion, the 9963, then on its alphabetics. Printed, the laboratory names precede the names ordered on alphabetics.

Examples of typical nomenclature index ordering compiled and printed by computer using the Nomenclature Sort Key Program are shown in Figures 5 through 7.



REGISTRY NUMBER	ITEM NUMBER	NAME TYPE	NAME, MOLECULAR FORMULA, AND SOURCES
137428	36	3	Carbathione $NS_2C_2H_5 \cdot Na$ CA
1327793	447	3	Carbazene Blue B CI
298464	8F	3	Carbazepine $N_2OC_{15}H_{12}$ CBAC
3240208	18	1	Carbazic acid, 3-(α-methylbenzyl)-, ethyl ester $N_2O_2C_{11}H_{16}$
69818	33	3	Carbazochrome $N_4O_3C_{10}H_{12}$ CAS
1428724	104	3	Carbazochrome sodium sulfonate $N_4O_5SC_{10}H_{11} \cdot Na$ INN,CBAC
86748	23	1	Carbazole $NC_{12}H_9$
	136	3	Carbazole $NC_{12}H_9$ CAS,CA,RI,CBAC,MERCK
6033881	11	1	Carbazole, picrate $NC_{12}H_9 \cdot xN_3O_7C_6H_3$ MERCK
6033870	18	1	Carbazole, potassium deriv $NC_{12}H_8 \cdot K$
	3A	3	Carbazole, N-potassium salt $NC_{12}H_8 \cdot K$ MERCK
6377124	37	1	Carbazole, 3-amino- $N_2C_{12}H_{10}$ CA,CI
132321	69	1	Carbazole, 3-amino-9-ethyl- $N_2C_{14}H_{14}$ CI
6402137	11	1	Carbazole, 2,7-diamino- $N_3C_{12}H_{11}$ CI
3244540	13	1	Carbazole, 3,6-dinitro- $N_3O_4C_{12}H_7$ CA,CI
86282	19	1	Carbazole, 9-ethyl- $NC_{14}H_{13}$ CA,CI
86204	15	1	Carbazole, 9-ethyl-3-nitro- $N_2O_2C_{14}H_{12}$ CI
1484124	19	1	Carbazole, 9-methyl- $NC_{13}H_{11}$ CA
6033905	12	1	Carbazole, 9-(phenoxyacetyl)- $NO_2C_{20}H_{15}$
6202159	15	1	Carbazole, 1,2,3,4-tetranitro- $N_5O_8C_{12}H_5$
1484135	12	1	Carbazole, 9-vinyl- $NC_{14}H_{11}$ CA
	205	3	Carbazole, 9-vinyl- $NC_{14}H_{11}$ CA
524801	12	1	Carbazole-9-acetic acid $NO_2C_{14}H_{11}$
6209230	13	1	Carbazole-9-acetic acid, ethyl ester $NO_2C_{16}H_{15}$
524801	45	3	9-Carbazoleacetic acid $NO_2C_{14}H_{11}$ MERCK
6209230	35	3	9-Carbazoleacetic acid, ethyl ester $NO_2C_{16}H_{15}$ MERCK
6407836	44	1	Carbazole-3-carboxanilide, 4'-chloro-1-[(2,4-dichlorophenyl)azo]-2-hydroxy- $N_4O_2Cl_3C_{25}H_{15}$
132616	16	1	Carbazole-3-carboxanilide, 4'-chloro-2-hydroxy- $N_2O_2ClC_{19}H_{13}$ CA
6407847	47	1	Carbazole-3-carboxanilide, 4'-chloro-2-hydroxy-1-[[2-methoxy-5-(phenylcarbamoyl)phenyl]azo]- $N_5O_4ClC_{33}H_{24}$
5894144	1A	1	Carbazole-9-propionitrile, 3-amino- $N_3C_{15}H_{13}$ CA
1327793	436	3	Carbazol Fast Blue CI
6411467	39	3	Carbazol Yellow $N_5O_8C_{26}H_{17} \cdot 2Na$ CI
86726	8D	4	p-(Carbazol-3-ylamino)phenol $N_2OC_{18}H_{14}$
	9E	3	p-(3-Carbazolylamino)phenol $N_2OC_{18}H_{14}$
538625	48	3	Carbazone, diphenyl- $N_4OC_{13}H_{12}$ CAS
60106	3C	3	Carbazone, diphenylthio- $N_4SC_{13}H_{12}$ CAS
88891	129	3	Carbazotic acid $N_3O_7C_6H_3$ MERCK,DPIM
524801	34	3	Carbazyl-N-acetic acid $NO_2C_{14}H_{11}$ MERCK
3240208	3A	3	Carbenzide $N_2O_2C_{11}H_{16}$ INN

Figure 5: Illustration of Ordering on Compound Parent



REGISTRY NUMBER	ITEM NUMBER	NAME TYPE	NAME, MOLECULAR FORMULA, AND SOURCES
75649	55	3	Butylamine, tertiary NC_4H_{11} MERCK
3037727	17	1	Butylamine, 4-(diethoxymethylsilyl)- $NO_2SiC_9H_{23}$ CA
4427763	53	3	Butylamine, 4,4'-iminobis- $N_3C_8H_{21}$ CA,CAS
110689	15	1	Butylamine, N-methyl- NC_5H_{13} CA
107857	38	3	Butylamine, 3-methyl- NC_5H_{13} CAS
109739	4C	3	n-Butylamine NC_4H_{11} CBAC,MERCK,DPIH
513495	13	1	sec-Butylamine NC_4H_{11} CAS,CA
75649	22	1	tert-Butylamine NC_4H_{11} CA
	77	3	tert-Butylamine NC_4H_{11} CAS
3858784	3A	3	n-Butylamine hydrochloride $NC_4H_{11}\cdot ClH$ CBAC
3785215	6C	3	ω-n-Butylaminoacetic acid 2-methyl-6-chloroanilide $N_2OClC_{13}H_{19}$ MERCK
	7D	3	1-(Butylaminoacetylamino)-2-chloro-6-methylbenzene $N_2OClC_{13}H_{19}$ MERCK
94257	37	3	Butyl amino benzoate $NO_2C_{11}H_{15}$ IP,MERCK,NF,ADI
	8C	3	Butyl p-aminobenzoate $NO_2C_{11}H_{15}$ CA,DCC,IECHTN
	6A	3	Butyl 4-aminobenzoate $NO_2C_{11}H_{15}$ IP
	48	3	n-Butyl p-aminobenzoate $NO_2C_{11}H_{15}$ NF,ADI,MERCK
5787268	34	3	Butyl aminobenzoate hydrochloride $NO_2C_{11}H_{15}\cdot ClH$ MERCK
577480	34	3	Butyl aminobenzoate picrate $NO_2C_{11}H_{15}\cdot {}^1/_2N_3O_7C_6H_3$ MERCK
	45	3	n-Butyl p-aminobenzoate picrate $NO_2C_{11}H_{15}\cdot {}^1/_2N_3O_7C_6H_3$ MERCK
94246	25A	4	p-(Butylamino)benzoic acid, 2-(dimethylamino)ethyl ester $N_2O_2C_{15}H_{24}$
1126789	47	3	Butylaniline $NC_{10}H_{15}$
	36	3	N-Butylaniline $NC_{10}H_{15}$ CAS
121006	4A	3	Butylatedhydroxyanisole $O_2C_{11}H_{16}$ CBAC
128370	103	3	Butylated hydroxytoluene $OC_{15}H_{24}$ VBB,CA,MERCK,CAS,CBAC,CFR
104518	5C	4	Butylbenzene $C_{10}H_{14}$
	4B	3	n-Butylbenzene $C_{10}H_{14}$ CBAC,MERCK
135988	35	3	sec-Butylbenzene $C_{10}H_{14}$ MERCK,CAS
5787291	33	3	(-)-sec-Butylbenzene $C_{10}H_{14}$ MERCK
5787280	3A	3	(+) sec-Butyl benzene $C_{10}H_{14}$ MERCK
98066	6B	3	tert-Butylbenzene $C_{10}H_{14}$ MERCK,CAS
5893834	48	3	2-[p-(5-tert-Butyl-2-benzimidozolylthio)phenylthio]-7-methylbenzothiazole $N_2S_2C_{25}H_{22}$ CA
136607	33	3	Butyl benzoate $O_2C_{11}H_{14}$ CAS,PCS
120503	59	3	Iso-Butyl benzoate $O_2C_{11}H_{14}$ CAS
571982	37	3	Butyl o-benzoylbenzoate $O_3C_{18}H_{18}$ IECHTN,CAS
129748	8H	3	1-(p-tert-Butylbenzyl)-4-(p-chlorodiphenylmethyl)piperazine dihydrochloride $N_2ClC_{28}H_{33}\cdot 2ClH$ MERCK
82951	108	3	1-(p-tert-Butylbenzyl)-4-(p-chloro-α-phenylbenzyl)piperazine $N_2ClC_{28}H_{33}$ INN

Figure 6: Illustration of Ordering Ignoring Prefixes

REGISTRY NUMBER	ITEM NUMBER	NAME TYPE	NAME, MOLECULAR FORMULA, AND SOURCES		
2429701	17	1	C.I. Direct Red 10	$N_5O_7S_2C_{32}H_{23}\cdot 2Na$	
2769075	12	1	C.I. Direct Red 17	$N_5O_7S_2C_{32}H_{23}\cdot 2Na$	
6772287	39	1	C.I. Direct Red 49	$N_8O_{13}S_4C_{35}H_{30}\cdot 4Na$	
6227083	34	1	C.I. Direct Red 55	$N_7O_{12}S_3C_{34}H_{25}\cdot 3Na$	
6470311	14	1	C.I. Direct Red 61	$N_6O_8S_2Cl_2C_{32}H_{22}\cdot 2Na$	
5915582	18	1	C.I. Direct Red 69	$N_4O_7S_4C_{31}H_{20}\cdot 2Na$	CI
6690706	11	1	C.I. Direct Red 83	$N_6O_{17}S_4C_{33}H_{24}\cdot 2Cu\cdot 4Na$	
	9A	3	C.I. Direct Red 121	$N_6O_8S_2C_{30}H_{24}\cdot 2Na$	CI
6470231	14A	1	C.I. Direct Red 123	$N_4O_6SC_{24}H_{20}\cdot Na$	
5938841	12D	1	C.I. Direct Red 127	$N_6O_8S_2C_{31}H_{26}\cdot 2Na$	
5873290	6A	1	C.I. Direct Red 127A	$N_6O_8S_2C_{31}H_{26}\cdot 2Na$	
	7B	3	C.I. Direct Red 127A	$N_6O_8S_2C_{31}H_{26}\cdot 2Na$	CI
5915628	48	1	C.I. Direct Red 130	$N_6O_{11}S_3C_{35}H_{28}\cdot 3Na$	
6417283	26	1	C.I. Direct Red 147	$N_6O_9S_2C_{30}H_{24}\cdot 2Na$	
6420399	182	1	C.I. Direct Red 149	$N_7O_8S_2C_{32}H_{25}\cdot 2Na$	
	17B	3	C.I. Direct Red 149	$N_7O_8S_2C_{32}H_{25}\cdot 2Na$	CI
5905226	9E	1	C.I. Direct Red 152	$N_7O_9S_2C_{36}H_{27}\cdot 2Na$	
	107	3	C.I. Direct Red 152	$N_7O_9S_2C_{36}H_{27}\cdot 2Na$	CI
5979317	21D	1	C.I. Direct Red 153	[illegible]	
6902256	4A	1	C.I. Direct Red 180	$N_{10}O_6C_{36}H_{32}\cdot 2Na$	
6406366	3A	1	C.I. Direct Red 186	$N_4O_9SC_{24}H_{16}\cdot 3Na$	
5905317	3A	1	C.I. Direct Red 189	$N_6O_{12}S_3C_{35}H_{28}\cdot 3Na$	
2586609	16	1	C.I. Direct Violet 1	$N_6O_8S_2C_{32}H_{24}\cdot 2Na$	
	469	3	C.I. Direct Violet 1	$N_6O_8S_2C_{32}H_{24}\cdot 2Na$	CA,CAS
6507831	12	1	C.I. Direct Violet 3	$N_4O_{11}S_3C_{32}H_{22}\cdot 3Na$	
	136	3	C.I. Direct Violet 3	$N_4O_{11}S_3C_{32}H_{22}\cdot 3Na$	CI
6227016	11	1	C.I. Direct Violet 5	$N_5O_7S_2C_{27}H_{21}\cdot 2Na$	CI
6227107	13	1	C.I. Direct Violet 7	$N_5O_7S_2C_{32}H_{29}\cdot 2Na$	CI
6417261	5D	1	C.I. Direct Violet 8	$N_4O_9SC_{32}H_{26}\cdot 2Na$	
	6E	3	C.I. Direct Violet 8	$N_4O_9SC_{32}H_{26}\cdot 2Na$	CI
6227141	15	1	C.I. Direct Violet 9	$N_5O_8S_2C_{30}H_{25}\cdot 2Na$	CI
6227196	1A	1	C.I. Direct Violet 11	$N_5O_8S_2C_{34}H_{27}\cdot 2Na$	CI
2429756	12	1	C.I. Direct Violet 12	$N_6O_8S_2C_{32}H_{24}\cdot 2Na$	
	329	3	C.I. Direct Violet 12	$N_6O_8S_2C_{32}H_{24}\cdot 2Na$	CI,CAS,CA
6227072	19	1	C.I. Direct Violet 16	$N_5O_{10}S_3C_{24}H_{21}\cdot 3Na$	CI
6470457	11	1	C.I. Direct Violet 21	$N_5O_7S_2C_{34}H_{27}\cdot 2Na$	
	99	3	C.I. Direct Violet 21	$N_5O_7S_2C_{34}H_{27}\cdot 2Na$	CI
6227118	16	1	C.I. Direct Violet 26	$N_5O_{10}S_3C_{34}H_{27}\cdot 3Na$	CI
6420060	19	1	C.I. Direct Violet 28	$N_4O_8S_2C_{34}H_{26}\cdot 2Na$	CAS,CI
6227130	12	1	C.I. Direct Violet 31	$N_5O_8S_2C_{30}H_{25}\cdot 2Na$	CI
6428940	69	1	C.I. Direct Violet 32	$N_5O_9S_2C_{34}H_{27}\cdot 2Na$	
	58	3	C.I. Direct Violet 32	$N_5O_9S_2C_{34}H_{27}\cdot 2Na$	CI
6227209	4B	1	C.I. Direct Violet 35	$N_5O_{11}S_3C_{34}H_{27}\cdot 3Na$	
6426773	3B	1	C.I. Direct Violet 38	$N_6O_{12}S_2C_{36}H_{28}\cdot 4Na$	
6059343	14	1	C.I. Direct Violet 39	$N_4O_8S_2C_{34}H_{26}\cdot 2Na$	
	69	3	C.I. Direct Violet 39	$N_4O_8S_2C_{34}H_{26}\cdot 2Na$	CA,CI
6227174	14	1	C.I. Direct Violet 40	$N_5O_8S_2C_{31}H_{27}\cdot 2Na$	CI
6369284	46	1	C.I. Direct Violet 41	$N_5O_7S_2C_{31}H_{27}\cdot 2Na$	
	57	3	C.I. Direct Violet 41	$N_5O_7S_2C_{31}H_{27}\cdot 2Na$	CI

Figure 7: Illustration of Ordering by Numeric Value When Alphabetics Are Identical

VII. REDESIGN OF THE REGISTRY COMPUTER SYSTEMS AND REPROGRAMMING FOR THE IBM 360 COMPUTERS

One of the most fundamental projects in Task I was started at the beginning of 1967 in order to completely redesign the Registry System for more efficient operation and to reprogram it for the IBM 360 computers. At the time, the Registry was operating on the IBM 7010 computers, and the basic programs had become inefficient due to the high level of modifications added to them as a result of initial operations. The 360 computer equipment represented a "third generation" of hardware capabilities and offered significantly improved processing times, processing capabilities, and capacity for future growth without major reprogramming.

As a result of the conversion to 360 equipment, CAS was able to develop and implement improved installation-wide standards for computer programs, to incorporate in the redesigned programs the results of experience with the first large-scale Registry operations, and to significantly improve the capacity for interface with other information processors and with future users of the System.

Among the overall improvements made in all Registry processing as a result of the conversion to 360 hardware are standard file formats, modular programming, and improved hardware capabilities. These are briefly summarized below, and are more fully described in Appendix F.

Standard file format refers to a CAS installation-wide standard for controlling the definition and use of each element of data used in any system. This format simplifies programming, assures consistency, and makes input, output, and search of the files easier.

Modular programming refers to the design of a computer system as a series of "pieces" (modules), each designed for a specific function. Modular programming increases the flexibility of a system by permitting modules to be changed individually without affecting other system operations.

Hardware improvements offered by the 360 include faster core access times, the ability to process data one-half bite at a time (offering a chance for compaction of the file), and the future potential for multiprogramming, direct access, and other advanced processing techniques.

A. Reprogramming the Structure Registry

The 360 Registry System was installed in April 1968 and has operated since that time. The technical improvements made in structure processing in this new System are described in Appendix F and are briefly indicated below.

1. Improved Handling of Tautomers

The new System automatically recognizes the equivalency of those unique compounds (called tautomers) for which two or more different but equally valid structural diagrams can be drawn. These diagrams do not represent isolable chemical species, and thus a single Registry Number is assigned to the tautomer no matter which alternative structure is entered. At the time the Registry Structure File was converted from 7010 format to 360 format, any tautomers on file were changed to the new representation.

2. Improved Handling of Rings

The redesigned registry programs incorporate new procedures for tracing paths in ring-containing compounds. The programs identify pairs of "ring closure" atoms as the starting point for the tracing of rings. This process is more efficient than the previous system of tracing rings from an arbitrarily selected atom.

3. Additional Editing Features

The new programs provide automatic editing and verification for certain established text descriptors, for Stock numbers (representing the oxidation states for selected multivalent metals), for "abnormal" atomic mass citations, and for the coordination numbers of coordination compounds.

4. Registration of Large Molecules

As a result of increased capacity in the 360 computers, certain arbitrary restrictions on the size and complexity of registered compounds were removed. The new system raises from 150 to 253 the number of nonhydrogen atoms accepted, and from six to 15 the number of nonhydrogen attachments allowed for any one atom. (See Figure 8.)

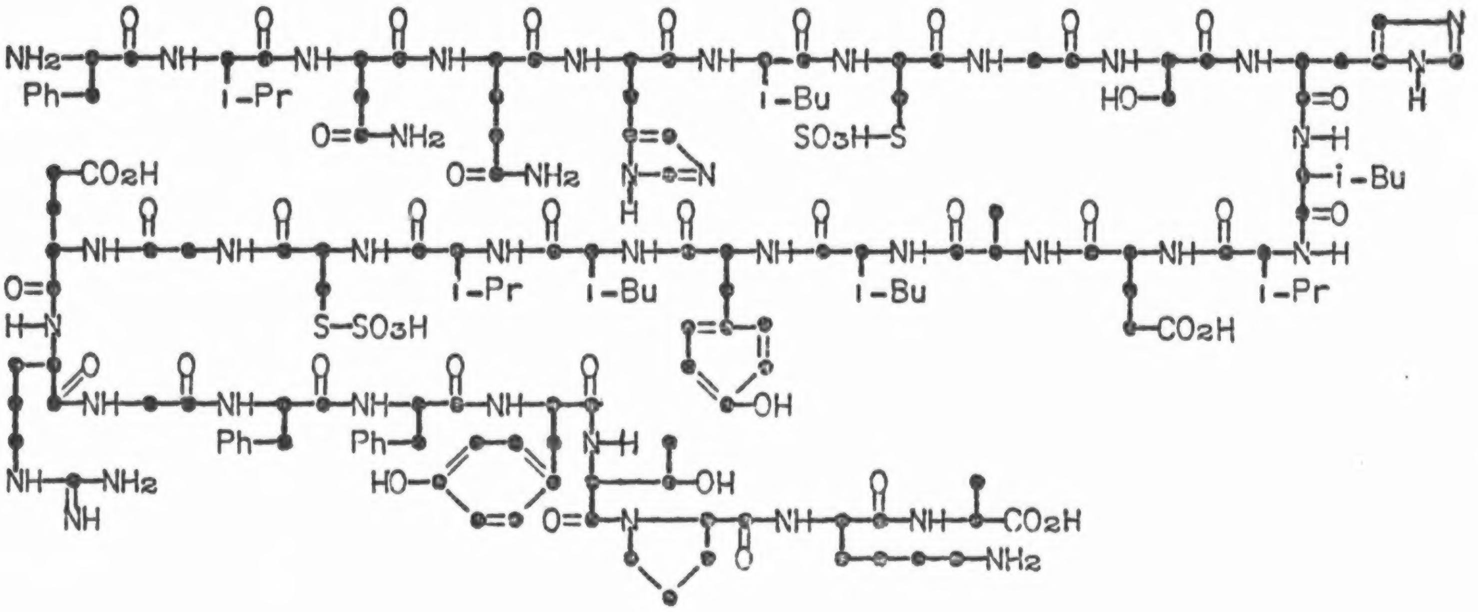
5. Structure Match without Registration

The new system permits a structure to be searched for in the Registry Files without adding it to the files if it is new. This feature will increase the file's usefulness to users with confidential or proprietary compounds who merely want to determine whether the compounds have yet been reported in the literature.

B. Reprogramming of Nonstructural Systems

The Registry's Bibliography and Nomenclature Systems were, in their 7010 versions, merely file maintenance systems with minimum output capability

```
®0018175605,foooz,ds,c157h232n40047s4,all.|
```



NH2
Ph
NH
i-Pr
O=C-NH2
i-Bu
SO3H-S
HO
N
H
O
i-Bu
CO2H
S-SO3H
OH
NH-C-NH2
NH
HO
CO2H
NH2

Figure 8: Typed structure input for the B chain of Bovine Insulin ($C_{157}H_{232}N_{40}O_{47}S_4$), a structure containing 248 nonhydrogen atoms. This example illustrates a large structure for which machine registration is possible in the 360 system, but was impossible in the 7010 system.

and with little capability for input editing and/or verification of names and references. In redesigning these systems for the IBM 360 computers, therefore, CAS had the objectives of greatly improved input editing capability and improved ability for flexible retrieval and file updating.

The redesigned systems included computer checking routines to support the keyboarding of names. These routines provide capitalization and italicization automatically, check, and in some cases correct, punctuation errors, and intercompare names to detect "problem" situations such as two names for a compound that differ only in locants, or two different CA Preferred Index Names for a compound. These improvements have been described in Section VI above.

The redesigned systems also incorporated output functions to maintain subfiles of the Registry System and create tape and printed listings of information according to various selection criteria imposed at output time.

Due to the complexity of the redesigned systems, they were first tested on a large subfile of the Registry System, the Common Data Base of Task III of Contract C414.* This large-scale test identified several problems, primarily in those aspects of the system that dealt with the retrieval and output functions. Therefore, CAS undertook to redesign these so-called "interface" functions of the Registry before making the new features generally available. The CAS Interface System is scheduled for implementation in the Autumn of 1969.

---

* See Final Report to the NSF on Contract NSF-C414, Task III, March 1969.

VIII. GLOSSARY

ACS

American Chemical Society

ALGORITHM

A precise step-by-step procedure which, when exactly followed, will result in a successful conclusion. No intellectual judgment is required in executing an algorithm.

ALTERNATING BONDS

A succession of alternating single and double bonds within a closed (cyclic) system.

BEILSTEIN

The term commonly used by chemists to refer to Beilsteins Handbuch der Organischen Chemie.

CA

Chemical Abstracts, the main publication of the Chemical Abstracts Service.

CAS

The Chemical Abstracts Service, a division of the American Chemical Society.

CAS CHEMICAL REGISTRY SYSTEM

An interrelated set of professional, clerical and computer-based processes that accomplish the registration of chemical compounds and the maintenance of information files resulting from the registration process. These files include compound records, molecular formulas, nomenclature, and bibliographic data. Registration of a compound involves the determination of the existence or nonexistence of a structural graph representation in the Registry Structure Files. The process includes the assignment of a unique number (Registry Number) to each substance that is new to the files.

CHARACTER SET

A specified selection of symbols (characters). The conventional computer printer character set consists of 48 upper case Latin letters, 10 Arabic numbers, and 12 symbols. CAS has a modified computer print chain which allows the use of a 120 character set. Each CA publication has specific character set requirements that are subsets of the CAS "universal character set" allowing for the upper and lower case Latin and Roman alphabet, Arabic number, superior and inferior characters, italic, lightface, and boldface characters, and a wide range of other symbols.

CHECK CHARACTER (check letter, check digits)

A technique used to verify the accuracy of recorded data. The check character is computed from the data content via an algorithm and is attached (usually suffixed) to the original data. Checking is accomplished by recomputing the check character and comparing it to the recorded character.

CHEMICAL COMPOUND

See COMPOUND.

CHEMICAL REGISTRY

See CAS CHEMICAL REGISTRY SYSTEM.

COLLECTIVE INDEX

One of a series of indexes, which, since the 5th Collective (1947 - 1956), have covered five years of CA. Previous Collectives covered 10 years. Each Collective Index combines the contents of the corresponding volume indexes.

COLLECTIVE PERIOD

The period during which material is accumulated for the next collective index. In 1967, CAS began the 8th Collective period, during which material will accumulate for the Eighth Collective Index (1967-71).

COMMON DATA BASE

This is made up of the published data associated with the substances specified in Task III for processing under this contract. The data elements in the Common Data Base are: (1) whatever structural description of the substance is provided; (2) Inverted Molecular Formula, when available; (3) Nomenclature; (4) Source Codes; (5) CA references for those synonyms taken from the Registry files.

COMPOUND

A single substance made up of identical molecular species.

COMPUTER-BASED

A computer-based process can be wholly computer executed, a computer assisted manual operation, or a process in which the results of a manual effort are subsequently recorded for computer processing.

CONNECTION TABLE

An atom-by-atom inventory of a molecule which shows each atom, the atoms connected to it, and the types of linkages (bonds). Mass number, coordination number, valences, and charges are shown wherever they are required for exact identification. Stereochemical data are included.

DATA AND INFORMATION

These terms are used interchangeably.

DEFINITIVE OLIGOMER

A compound whose structure is represented by that of a single unit of known structure repeated a known number of times. The structure of the definitive oligomer is known only to the extent that it is a multiple of the single unit.

DESKTOP ANALYSIS TOOLS

Printed lists of specially selected substances for use in the analysis of input information to help identify already registered substances. These tools include listings of names, laboratory numbers, and acronyms with the associated Registry Numbers, molecular formulas and source codes.

FLAGGING

Flagging is a process of including control symbols in the stream of recorded information to indicate special conditions to the computer program.

FORMULA INDEX

A compilation of molecular formulas arranged in order. Specifically, one of the published indexes to CA.

GRAPH (of a structure)

The basic pattern of nonhydrogen nodes and their connecting lines in a structural diagram, with no designation as to the specific elements or types of bonds present.

8

HARD COPY

A printed copy of machine output in a readable form for human beings, e.g., printed reports, listings, documents, summaries, etc.

HARDWARE

Physical equipment.

HIERARCHY

A series of objects, concepts or indexing terms divided or classified in ranks or orders, as in a family tree or a botanical classification. A genus-species relationship is a particular type of hierarchical relationship.

KEYBOARD

verb: To record data on an information storage medium (e.g., papers, magnetic tape, punched cards, etc.) using a finger-operated set of keys.

noun: The assemblage of finger-operated keys used on an information-recording machine such as a typewriter, keypunch, adding machine, etc.

LIGAND

The group, in coordination compounds, which either: (1) is the donor atom; or (2) contains the donor atom(s).

LINE

The site of a bond in a structural diagram.

MACHINE-READABLE

Refers to any representation of data in a form directly acceptable to a machine, specifically, to a computer. Punched cards, punched paper tape, and magnetic tape, for example, may all contain information in machine-readable form.

MAGNETIC TAPE

A plastic tape impregnated or coated with magnetic material on which information may be recorded by computer or keyboard-driven devices. Magnetic tape is one of several means to store information for subsequent reprocessing by computer. Excluded from the usage of this term in this document is magnetic tape used for sound or video recording.

MANUAL

Refers to processes handled largely by human effort.

MANUAL REGISTRATION

A system of registration in which a chemist, working with a small set of files, determines the uniqueness of a substance and, on the basis of this determination, assigns the substance a Registry Number or retrieves the one previously assigned. Manual registration is used for a small minority of substances that do not meet the criteria for machine registration.

MECHANICAL REGISTRATION

A computer-based system of registration.

MECHANIZED (mechanized)

For purposes of this document: synonymous with computer-based.

MODULAR

In computer systems design, modular systems have the following attributes:

The total entity system is broken into explicitly defined sub-units (Modules).

Each sub-unit is functionally independent of the other sub-units.

Modules have standard points and methods of interfacing with other modules.

Modification to a module does not affect other modules if only the activity is changed and not the interface.

MOLECULAR FORMULA

A listing of each kind of element and total number of atoms of each kind present in a molecule.

MOLFORM

Molecular Formula.

MONOMER

The unpolymerized form of a compound that can be polymerized.

MULTIPROGRAMMING

A computer processing technique that provides for the operation of more than one application program in a computer system at the same time. Program execution alternates among the individual programs.

NAME MATCH

As a registration process, a technique for determining identical compounds by comparing names in order to locate compounds and related data which might be identical.

NODE

A word used interchangeably with "atom" in the context of structural diagrams and connection tables.

NOMENCLATURE

All types of names used to identify chemical compounds including acronyms and laboratory numbers. Preferred nomenclature refers to the compound names preferred for use in current CA Subject and Formula Indexes. Systematic nomenclature are any names that are systematically derived from the structure of the chemical substance.

ORGANIC COMPOUNDS

For the purposes of this report these are carbon-containing compounds exclusive of coordination complexes.

PAPER TAPE

A tape on which a pattern of holes or cuts is used to represent data.

POLYMER

A compound or mixture of compounds each of which consists essentially of repeating structural units.

PROGRAM

The complete sequence of machine instructions and routines necessary to solve a problem.

PUNCHED CARD

A card of lightweight cardboard on which information is represented by holes punched in specific positions, which may be processed by automatic machinery, semi-automatically, or manually.

RECORDING

An interrelated set of activities that cause information to be stored in mechanized files or on a printed page. Recording includes transcription or dictation, keyboarding, and entry to the computer files.

REGISTRATION

The process of determining the existence or nonexistence of a substance in the Registry Files. The process includes the assignment of a unique number (Registry Number) to each substance that is new to the files; this number is used in a large, multifaceted system to associate data related to that substance.

REGISTRY FORM

A data and worksheet especially designed for Registry System use to contain all data needed for Registry processing and including data to be filed in the Registry. Included are the structure drawing, molecular formula, CA reference, author name for the compound, connection table, etc. In general, one compound will be represented by one Registry Form.

REGISTRY NUMBER

The unique number which is assigned to each substance when it first enters the Registry and which is recalled each time that substance is checked against the file. The Registry Number may be used to identify fully the substance, and in the future it can be used as the address in specialized subject files to identify data associated with the substance. A Registry Number may include alphabetic characters, and will include a computed check digit.

RING

A group of atoms and their bonds which form a closed loop.

RING CLOSURE PAIRS

A field of the structural record which lists pairs of atoms with a connecting bond which constitute the uniting of the ends of a string of atoms forming a ring.

SOURCE CODE

A set of codes attached to each name in the Registry Nomenclature File that identifies:

1. Source(s) in which the name is used. These may be published sources, for example Journal of Biological Chemistry, Chemical Abstracts, and Merck Index, or private sources, for example a file belonging to a given organization.
2. Organizations which have some association with the substance and/or its name.

STEREOCHEMICAL DESCRIPTOR

An abbreviated form of the traditional designation of stereochemistry.

STRUCTURE

The mode of linkage of atoms in a chemical molecule.

STRUCTURAL DIAGRAM

A two-dimensional graphic representation of the atoms and bonds of a molecule.

SUBSTRUCTURE

A specified set of atoms interconnected in a specified way. This constellation normally represents less than a complete molecule.

SUBSTRUCTURE SEARCH

The search through a file of representations of structures of compounds for a specified set of atoms connected in a specified way. The set normally represents less than a complete molecule.

SUBSTRUCTURE SEARCH SYSTEM

A computer-based structure retrieval system designed for generic searching of structural information (stored in the CAS Chemical Registry).

TAUTOMER END POINTS

A field of the structural record which describes the end points of a tautomer string in the graph proper. It is understood that the two atoms in each of these tautomer end point pairs are sharing one hydrogen atom.

TAUTOMERS

Two or more structures which are considered valid representations of a given substance.

TEXT DESCRIPTOR

A symbol, word,or words appended to the computer structural record which is used to describe the stereochemistry and/or other information connected with that structure which cannot be described adequately in the fields of the graph proper or in the other modification fields.

TID NUMBER

The Temporary Identification number used to identify a compound during the input steps prior to registration.

TRIVIAL NAME

A name or number which is not systematically derived by consideration of the structure of the corresponding compound.

VALENCE

The sum of the H-count, the numerical value of charges, and the value of bonds (1,2, or 3) to an atom. (For the purposes of registration and searching.)

VERIFY

The process of assuring that recorded material agrees with the edited manuscript. The process may or may not include literal verification as commonly used in conjunction with keypunching. Verification can also imply proofing material to assure the computer records are correct in terms of released manuscripts.

APPENDIX A

An Overview Description of the CAS Chemical Registry System

Extracted from Substructure Search "Background Information and Question Coding Instructions"

Copyright 1968 by the American Chemical Society

Taken From
Substructure Search
Background Information and Question Coding Instructions
Published by the American Chemical Society

Chapter 2: BACKGROUND INFORMATION ON THE CAS CHEMICAL COMPOUND REGISTRY SYSTEM

2.1 Introduction

The Chemical Compound Registry System* being implemented at CAS is a man-machine system for the storage and retrieval of chemical nomenclature, bibliographic data, and structural information. The basis for registration is a machine-derived notation unique to each compound and representing that compound's molecular structure in detail equivalent to that provided by the conventional structural diagram used as a communication tool by chemists. This chapter describes the Registry System with particular emphasis on those aspects of the system that influence Substructure Searching.

2.2 General Contents of the Registry Structure File

The Structure File of the Registry System presently contains current information on more than 1,000,000 substances, including those which have been indexed in Chemical Abstracts beginning with Volume 62 (1965), plus information from several other sources. All compounds having fully defined structures (with few exceptions — oligomers for example) are registered and are searchable in accordance with appropriate input structure-drawing conventions. The Registry includes organic compounds, inorganic compounds, and coordination compounds (compounds in which the number of attachments to a central atom — usually a metal atom, but not necessarily — exceeds the generally accepted valence of the central atom). Text descriptors are employed for differentiation of stereoisomers. Mixtures (with the exception of racemic mixtures) are not presently registered by structure and are therefore not retrievable by substructure search.

---

* D. P. Leiter, Jr., H. L. Morgan, R. E. Stobaugh, "Installation and Operation of a Registry for Chemical Compounds," J. Chem. Doc. **5** (4), 238 (1965).

One group of partially defined structures, the definitive oligomers such as dimers, trimers, etc., are registered by structuring the monomer according to standard Registry procedures and by using a text descriptor to indicate the number or repeating units. Complete polymer registration procedures are still under development and will be implemented at a later date.

Elements, their isotopes, and allotropes are registered in accordance with Chemical Abstracts indexing policies and are searchable as registered. Compounds having unknown or incompletely defined structures are not searchable by substructure since there are not entries in the Structure File. Methods for equalization of certain types of tautomers have also been incorporated.

2.3 The Computer Structural Record

The structure for each compound is stored in the form of a compact list* containing all non- H-atom connections, element symbols, bond values, H-counts for non-carbon atoms, ring closure pairs, and various modifications such as abnormal mass, charge, and abnormal valence. The format of this compact list, as translated from the tape version to a printed form,* is illustrated with annotations in Figure 2.1 for a compound regiring no special structuring conventions or machine manipulations to facilitate input. Those compound types requiring particular conventions for registration are discussed in Section 2.4 with examples and explanations of the resultant structural records.

---

* cf. footnote page 1.

```
              REG NO 76153
                TEXT NS


ATOM NO  1   2   3   4   5   6   7   8
CONN.        1   1   1   1   2   2   2
ELEMENT  C   C  CL   F   F   F   F   F
BOND        -1  -1  -1  -1  -1  -1  -1
H-COUNT
```

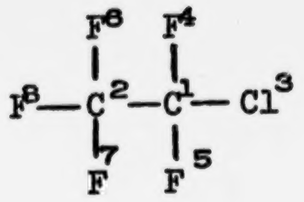


Figure 2.1: COMPACT LIST FOR A SIMPLE GRAPH

---

* A detailed description of the machine record format is available as part of the documentation of software for the Registry System.

Description of Fields (cf. Figure 2.1)

REG NO (Registry number) - This is the machine-assigned unique permanent address for a compound and its related information on all CAS Registry files.

TEXT - This field carries textual information qualifying the structural record of a compound (e.g., stereochemistry and repetition of structural units).

ATOM NO - This field designates the machine numbers assigned to the non-^{1}H atoms of the structure. The connection, element symbol, bond value, and ^{1}H-count (for non-carbon atoms) appear in their respective fields in the column below the atom number.

CONN. - This field carries the description of atom connections in the form of a "from list". It records the lowest numbered atom to which each atom in the ATOM NO field is attached (ring closures are handled separately).

ELEMENT - This field records the element symbol for each machine numbered atom in the column beneath the appropriate atom number. Any element except ^{1}H can appear in this field. (Deuterium "D" and tritium "T" are handled as separate elements.)

BOND - This field is the description of the bond between an atom and the atom to which it is connected, as designated in the CONN. field. Each bond description consists of two characters - the first character is either the operator "-" representing acyclic (chain) bonds, or the operator

"*" representing cyclic (ring) bonds. The second character can be a number from 1 to 5 which represents the following bond values:

1 = single
2 = double
3 = triple
4 = equalized tautomer (cyclic and acyclic)
5 = completely conjugated cyclic

H-COUNT - This field records the number of H atoms attached to non-carbon atoms of the structure. The H-count is placed below the atom number of the respective atoms.

With the fields outlined thus far, the compact list of the structure in Figure 2.1 would read as follows:

> Registry number 76,153 is the permanent address of the compound, and no stereochemistry (NS) has been described for the structure. There are 8 atoms in the structure: atom 1 is a carbon atom; atom 2 is a carbon atom connected to atom 1 by an acyclic single bond; atom 3 is a chlorine atom connected to atom 1 by an acyclic single bond; atom 4 is a fluorine atom connected to atom 1 by an acyclic single bond; atom 5 is a fluorine atom connected to atom 1 by an acyclic single bond; atom 6 is a fluorine atom connected to atom 2 by an acyclic single bond; atom 7 is a fluorine atom connected to atom 2 by an acyclic single bond; and atom 8 is a fluorine atom connected to atom 2 by an acyclic single bond.

Thus, all atoms, bonds, and connections are completely and uniquely described for the structure in Figure 2.1. For more complex structures, as illustrated in Figure 2.2, more fields are required to describe the

REG NO 5611518
TEXT 11B,16A-PREGN

ATOM NO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
CONN.		1	1	1	1	2	2	2	3	3	4	5	5	6	7	11	11	12	14	14
ELEMENT	C	C	C	O	C	C	C	C	C	O	C	C	O	C	C	C	C	O	C	C
BOND		*1	*1	*1	-1	*1	*1	-1	*1	*1	*1	-1	-2	*1	*1	-1	-1	-1	*1	*1
H-COUNT																				

ATOM NO	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
CONN.	15	18	19	19	20	22	22	23	23	23	26	28	29	31	31	31	32	37
ELEMENT	O	C	C	F	C	C	O	C	C	C	C	C	C	C	C	C	C	O
BOND	-1	-1	*1	-1	*1	-1	-2	*1	*1	-1	-1	*2	*2	-1	-1	-1	*1	-2
H-COUNT	1																	

RING CLOSURE PAIRS 006*1009 010*1011 015*1019 025*1028 033*1037

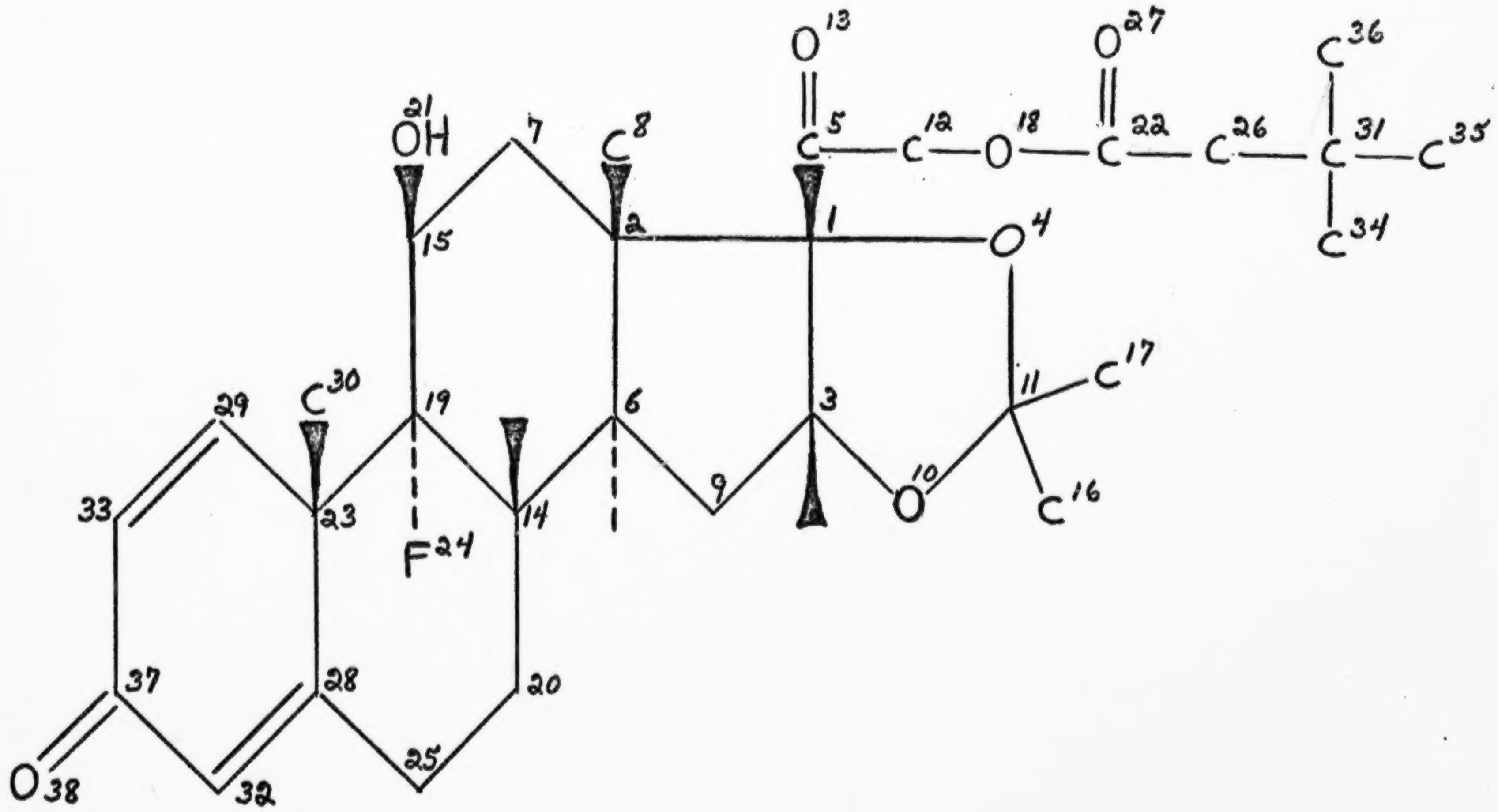

O 13
O 27
C 36
OH 21
7
C 8
C 5
C 12
O 18
C 22
C 26
C 31
C 35
15
2
1
O 4
C 34
C 17
11
C 30
29
19
6
3
33
23
F 24
14
9
O 10
C 16
37
28
20
25
O 38
32

Figure 2.2: COMPACT LIST FOR A STRUCTURE WITH RINGS

additional structural details. Figure 2.2 illustrates the treatment of a compound containing rings. The resultant compact list contains a field entitled RING CLOSURE PAIRS.

RING CLOSURE PAIRS - This eight-character subfield lists two three-character atom numbers with a two-character bond descriptor between them. This means that the two atoms cited are connected by the cyclic bond value listed. The ring closure list for the structure in Figure 2.2 would read as follows:

> Atom 6 is connected to atom 9 by a single cyclic bond; atom 10 is connected to atom 11 by a single cyclic bond; atom 15 is connected to atom 19 by a single cyclic bond; atom 25 is connected to atom 28 by a single cyclic bond; and atom 33 is connected to atom 37 by a single cyclic bond.

The fields ATOM NO, CONN., ELEMENT, BOND, H-COUNT, and RING CLOSURE PAIRS collectively are termed the "graph proper" of the structural record. All other fields are collectively termed "modifications" and are explained below:

AB VAL (Abnormal Valence) - Table 2.1 is a table of standard valences. For registration and searching, the valence of an atom is defined as the sum of the H-count, the numerical value of charges, and the value of bonds (1, 2, or 3) to that atom (e.g., in $H_3C-CH_2-N=\overset{\oplus}{N}=\overset{\ominus}{N}$, the atom valences left to right are C=4, C=4, N=3, N=5, N=3). The valences of atoms in a completely conjugated cyclic system (or a tautomeric string) can be calculated

Table 2.1: STANDARD VALENCES FOR THE ELEMENTS

Symbol	Element	Valence	Symbol	Element	Valence
Ac	Actinium	3	Dy	Dysprosium	3
Ag	Silver	1	Er	Erbium	3
Al	Aluminum	3	Es	Einsteinium	3
Am	Americium	3	Eu	Europium	3
Ar	Argon *	0	F	Fluorine	1
As	Arsenic	3	Fe	Iron	2
At	Astatine	1	Fm	Fermium	3
Au	Gold	1	Fr	Francium	1
B	Boron	3	Ga	Gallium	3
Ba	Barium	2	Gd	Gadolinium	3
Be	Beryllium	2	Ge	Germanium	4
Bi	Bismuth	3	H	Hydrogen	0
Bk	Berkelium	3	He	Helium *	0
Br	Bromine	1	Hf	Hafnium	4
C	Carbon	4	Hg	Mercury	2
Ca	Calcium	2	Ho	Holmium	3
Cd	Cadmium	2	I	Iodine	1
Ce	Cerium	3	In	Indium	3
Cf	Californium	3	Ir	Iridium	2
Cl	Chlorine	1	K	Potassium	1
Cm	Curium	3	Kr	Krypton *	0
Co	Cobalt	2	La	Lanthanum	3
Cr	Chromium	6	Li	Lithium	1
Cs	Cesium	1	Lu	Lutetium	3
Cu	Copper	2	Lw	Lawrencium	3
D	Deuterium	1	Md	Mendelevium	3
			Mg	Magnesium	2

---

* An oxidation number of zero (0) is used for the rare gases.

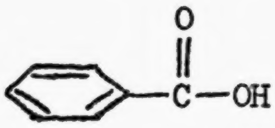
Table 2.1: (continued)

Symbol	Element	Valence	Symbol	Element	Valence
Mn	Manganese	7	Ru	Ruthenium	2
Mo	Molybdenum	6	S	Sulfur	2
N	Nitrogen	3	Sb	Antimony	3
Na	Sodium	1	Sc	Scandium	3
Nb	Niobium	5	Se	Selenium	2
Ne	Neon *	0	Si	Silicon	4
Nd	Neodymium	3	Sm	Samarium	3
Ni	Nickel	2	Sn	Tin	4
No	Nobelium	3	Sr	Strontium	2
Np	Neptunium	5	T	Tritium	1
O	Oxygen	2	Ta	Tantalum	5
Os	Osmium	2	Tb	Terbium	3
P	Phosphorus	3	Tc	Technetium	7
Pa	Protactinium	5	Te	Tellurium	2
Pb	Lead	4	Th	Thorium	4
Pd	Palladium	2	Ti	Titanium	4
Pm	Promethium	3	Tl	Thallium	1
Po	Polonium	2	Tm	Thulium	3
Pr	Praseodymium	3	U	Uranium	6
Pt	Platinum	2	V	Vanadium	5
Pu	Plutonium	4	W	Tungsten	6
Ra	Radium	2	Xe	Xenon	0
Rb	Rubidium	1	Y	Yttrium	3
Re	Rhenium	7	Yb	Ytterbium	3
Rh	Rhodium	2	Zn	Zinc	2
Rn	Radon *	0	Zr	Zirconium	4

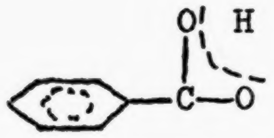
---

* An oxidation number of zero (0) is used for the rare gases.

by using one of the Kekule forms (as opposed to the machine equalized form) as the basis for determining bond values and ^{1}H count (e.g.,



instead of



would be used as the basis for calculating valences). Valences other than those given in Table 2.1 (Standard Valences for the Elements) are recorded as modifications in the field designated "AB VAL". For example, the abnormal valence field of Figure 2.3 contains the following: AB VAL 003-005 025-006. This would read: atom number 3 has a valence of 5, atom number 25 has a valence of 6, etc.

AB MASS (Abnormal Mass) - The structure registration process checks abnormal mass citations against a list of acceptable abnormal masses (Table 2.2). This list includes the abnormal masses most commonly cited in chemical texts, reference works, and abstracts.

Table 2.2: TABLE OF ACCEPTABLE ABNORMAL MASS VALUES

Element Symbol	Acceptable Mass Values
Au	195, 198, 199
Br	77, 79, 81, 82
C	11, 13, 14
Ca	45, 47
Cl	36, 38
Co	56, 57, 58, 60
Cr	51
I	124, 125, 129, 131, 132
K	42
N	15
Na	24
O	17, 18
P	32
S	35
Sr	90

REG NO 10380162
TEXT CIS

ATOM NO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
CONN.		1	1	2	3	3	3	4	5	8	10	11	12	13	14	15	16
ELEMENT	C	C	N	C	C	C	C	C	C	C	C	C	C	C	C	C	C
BOND		-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	-1
H-COUNT																	

ATOM NO	18	19	20	21	22	23	24	25	26	27	28	29	30
CONN.	17	18	19	20	21	22		24	24	25	25	25	26
ELEMENT	C	C	C	C	C	C	O	S	C	O	O	O	C
BOND	-1	-1	-1	-1	-1	-1		-1	-1	-1	-2	-2	-1
H-COUNT													

AB VAL 003-005 025-006
CHARGE 003,+1 027,-1
M.A.F. $\underbrace{024}_{\text{beginning atom of 2nd fragment}},\underbrace{01/001}_{\text{ratio to 1st fragment}}$

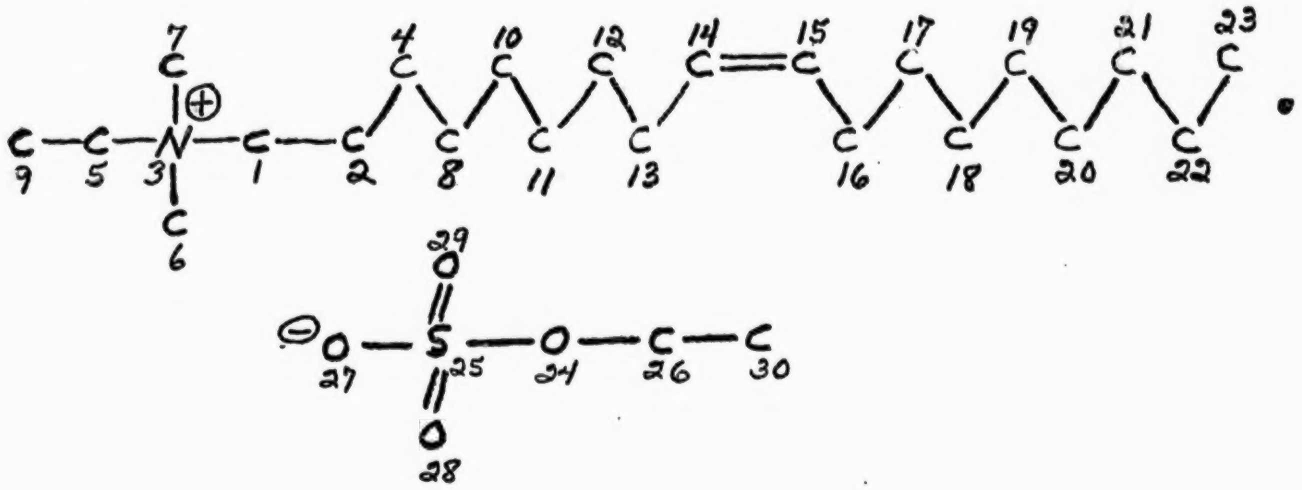

7
4
10
12
14
15
17
19
21
23
C
C
C
C
C
C
C
C
C
C
⊕
C
C
N
C
C
C
C
C
C
C
C
C
9
5
3
1
2
8
11
13
16
18
20
22
C
6
29
O
⊖O
S
O
C
C
27
25
24
26
30
O
28

Figure 2.3: COMPACT LIST FOR A GRAPH CONTAINING A MULTI-ATOM FRAGMENT

Acceptable abnormal masses are recorded as modifications in the field designated "AB MASS". For example, the abnormal mass field in Figure 2.4 contains the following: AB MASS 001-032. This would read: atom number 1 has a mass of 32.

CHARGE - Charges on atoms are recorded as modifications to the graph proper in a field designated "CHARGE". Charges in the range -9 through +9 are acceptable, and they must be linked to specific atoms in the machine record. (See conventions for coordination compounds in Section 2.4). For example, the charge field of Figure 2.3 contains the following: CHARGE 003,+1 027,-1. This would read: atom number 3 has a charge of +1, atom number 27 has a charge of -1, etc.

S.A.F. (Single-Atom-Fragment) - See dot-disconnected conventions in Section 2.4.

M.A.F. (Multi-Atom-Fragment) - See dot-disconnected conventions in Section 2.4.

TAUTOMER ENDPOINTS - See tautomer description in Section 2.5.

2.4 Structuring Conventions and Resultant Computer Records

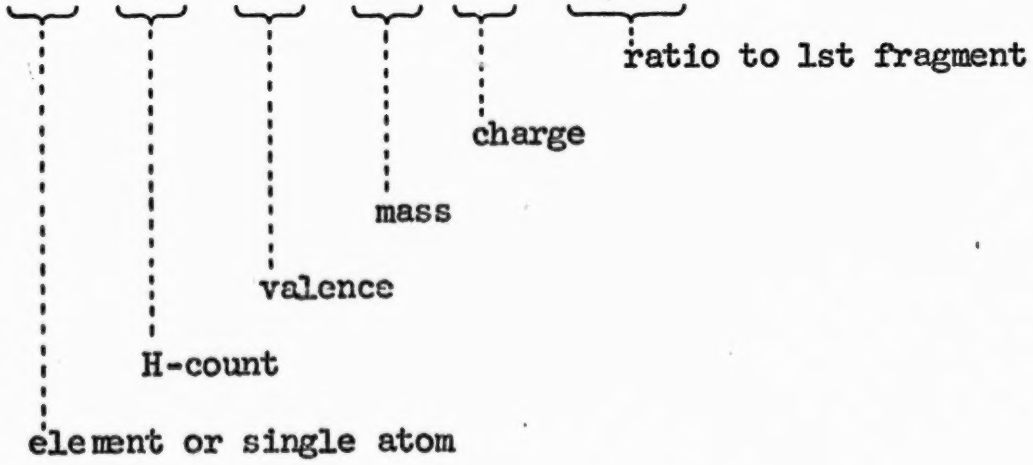
The structuring conventions employed to input structural diagrams must result in consistently generated compact lists to insure unique registration and reliable retrieval of structural information. This section provides information on those conventions which must be considered when phrasing substructure search questions.

REG NO 7635463
TEXT NS

ATOM NO	1	2	3	4	5
CONN.		1	1	1	1
ELEMENT	P	O	O	O	O
BOND		-4	-4	-4	-4
H-COUNT					

TAUTOMER ENDPOINTS 002-005, 003-005, 004-005

AB VAL 001-005
AB MASS 001-032
S.A.F. NA 000 001 000 02/001


ratio to 1st fragment
charge
mass
valence
H-count
element or single atom

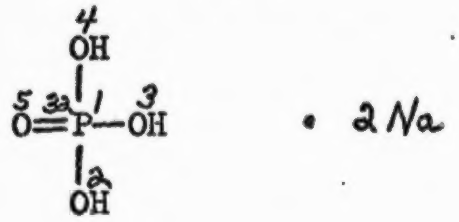

4
OH
5
3a
1
3
O
P
OH
2
OH
• 2Na

Figure 2.4: COMPACT LIST FOR A GRAPH CONTAINING A SINGLE-ATOM FRAGMENT

Table 2.3: REGISTRY REPRESENTATION FOR CERTAIN COMPOUNDS OR GROUPS

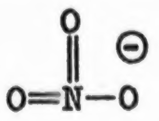
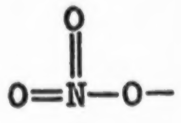
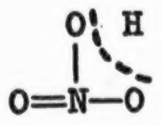
Name	Formula	Registry Representation
Azide	$N_3^{\ominus}$ (in NaN_3)	$\overset{\ominus}{N}=\overset{\oplus}{N}=\overset{\ominus}{N}$
Azido	$-N_3$ (in RN_3)	$-N=\overset{\oplus}{N}=\overset{\ominus}{N}$
Carbon monoxide	CO	$\overset{\oplus}{C}\equiv\overset{\ominus}{O}$
Diazo	$=N_2$ (in CH_2N_2)	$=\overset{\oplus}{N}=\overset{\ominus}{N}$
Diazonium	$-\overset{\oplus}{N_2}$ (in $R\overset{\oplus}{N_2}\,\overset{\ominus}{Cl}$)	$-\overset{\oplus}{N}\equiv N$
Fulminic acid	HON:C	$HO-\overset{\oplus}{N}\equiv\overset{\ominus}{C}$
Isonitrile	$-NC$ (in $R-NC$)	$-\overset{\oplus}{N}\equiv\overset{\ominus}{C}$
Nitrate	NO_3 (in $NO_3^{\ominus}$ or $RONO_2$)	 $O=\overset{\overset{O}{\Vert}}{N}-\overset{\ominus}{O}$ or  $O=\overset{\overset{O}{\Vert}}{N}-O-$
Nitric acid	HNO_3	

Table 2.3 (continued)

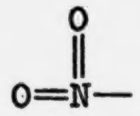
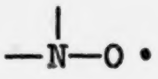
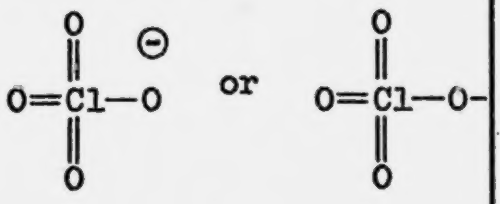
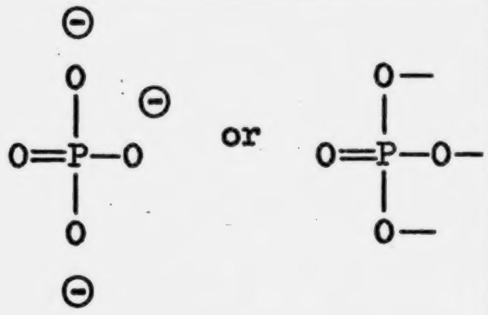
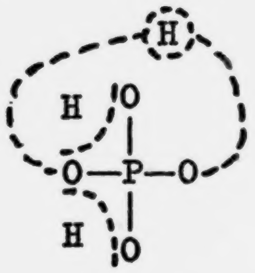
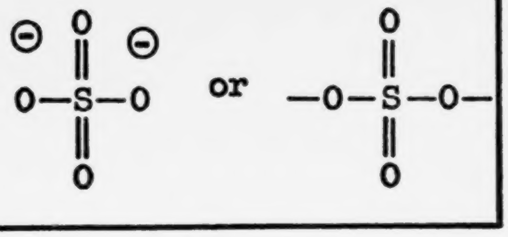
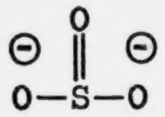
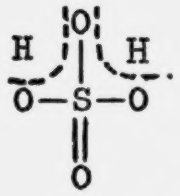
Nitrite	NO_2 (in $NO_2^{\ominus}$ or RONO)	$O{=}N{-}O^{\ominus}$ or $O{=}N{-}O$
Nitro	$-NO_2$	 $\overset{O}{\overset{\Vert}{O{=}N{-}}}$
Nitroxide	NO (in R_2NO)	 $-\overset{\vert}{N}-O\bullet$
Perchlorate	ClO_4 (in $ClO_4^{\ominus}$ or $ROClO_3$)	
Phosphate	PO_4 (in $PO_4^{\equiv}$ or R_3PO_4)	
Phosphoric acid	H_3PO_4	
Sulfate	SO_4 (in $SO_4^{=}$ or R_2SO_4)	

Table 2.3 (continued)

Sulfite	SO_3 (in $SO_3^{=}$ or R_2SO_3)	
Sulfuric acid	H_2SO_4	

Rigorously Defined Structures - Certain compounds and functional groups have been rigorously defined to insure consistent registration and retrieval. A number of these are listed in Table 2.3 with their registry representations.

Dot-Disconnected Conventions - Amine salts, hydrates, and π-complexes are, in general, handled as addition compounds. The interconnections between the structural formulas of the various components are shown by a dot (e.g., $CH_3NH_2 . H_2SO_4$). Metal salts of oxy acids (carboxylic,sulfonic,etc.) and metal derivatives of many other functional groups containing -OH, -SH, -SeH, -TeH, and -NH are handled by a similar dot-disconnected convention *. That is, the structural formula of the acid (or other compound) is followed in turn by a dot, a coefficient depicting the ratio of the first fragment, and the metal (e.g., $CH_3COOH \cdot \frac{1}{2}Ca$; $CH_3OH \cdot Na$).

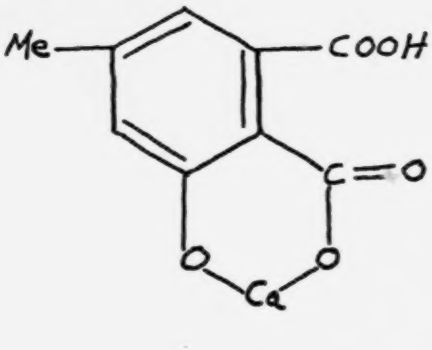
Table 2.4 contains the list of elements defined as nonmetals for the purposes of registration and searching.

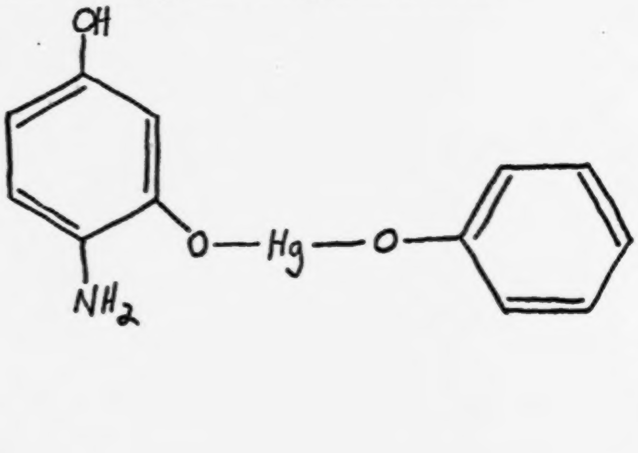
Table 2.4: LIST OF NONMETALS

Antimony
Argon
Arsenic
Astatine
Boron
Bromine
Carbon
Chlorine
Fluorine
Helium
Hydrogen
Iodine
Krypton
Neon
Nitrogen
Oxygen
Phosphorus
Radon
Selenium
Silicon
Sulfur
Tellurium
Xenon

---

* Except when a multivalent metal atom is in a ring or unsymmetrical structure as in:





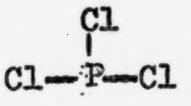
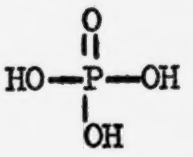
Metal salts or derivatives of organic compounds containing functional groups of the nonmetals other than N, O, S, Se, and Te are structured and stored using bond connections (i.e., they are not structured by the dot-disconnected convention). Metal derivatives of strictly inorganic hydrides of all nonmetals (including N, O, S, Se and Te) are registered manually and, therefore, are not searchable by structure (e.g., NaOH, $LiNH_2$)*.

M.A.F. (Multi-Atom-Fragment) - Figure 2.3 gives an example of a dot-disconnected structure in which both fragments contain more than one non-^{1}H-atom. Such fragments are recorded as part of the graph proper with discontinuities marking the separation of fragments. The ratios of multi-atom-fragments to the first fragment of the structure are entered as a modification in the Multi-Atom-Fragment (M.A.F.) field. Thus, the M.A.F. field of Figure 2.3 would read: the fragment beginning with atom 24 in the graph proper has a ratio of 1 to 1 with the first fragment of the graph proper.

S.A.F. (Single-Atom-Fragment) - Figure 2.4 gives an example of a compact list representing a dot-disconnected structure in which one portion of the structure is a single atom of the metal sodium. When one fragment of a dot-disconnected structure contains only one non-^{1}H atom or a non-^{1}H atom with attached H (e.g., H_2O, NH_3, and ^-OH), no indication of this fragment is given in the graph proper. Instead, a special Single-Atom-Fragment (S.A.F.) field is set up as a modification designating the single atom of the fragment with its ^{1}H-count, valence, abnormal mass, charge, and ratio to the first fragment of the structure. Thus, the S.A.F. field of Figure 2.4 would read: there is a single-atom-fragment consisting of Na with a H-count of 000, a valence of 001, a normal mass (indicated by 000 in this position), a charge of +0, and a ratio of 2 to 1 to the first fragment of the graph proper.

---

*Additional details on the treatment of metal salts and other metal derivatives are available in "Chemical Abstracts Service Chemical Compound Registry - Structure Conventions" January 31, 1968.

Inorganic Compounds - Those compounds containing no carbon in which the standard valence of each atom is filled, but not exceeded, are structured using the same covalent structuring conventions used for organic compounds (e.g.,  $\mathrm{Cl{-}\overset{\displaystyle Cl}{\overset{|}{P}}{-}Cl}$ and Na—Cl). Those inorganic compounds in which the number of connections to a central metal atom exceeds the value of the oxidation state ‡ for that atom are structured using the conventions for coordination compounds outlined below. Metal salts of inorganic oxo acids (and their S, Se and Te analogs) are structured using the dot-disconnected convention (e.g.,  $\mathrm{HO{-}\underset{\displaystyle OH}{\underset{|}{\overset{\displaystyle O}{\overset{\|}{P}}}}{-}OH \cdot 3/2\ Ca}$).

Coordination Compounds - The structuring conventions for metal coordination compounds are designed for consistency in structural representation and consistency in nomenclature. The system which has been adopted indicates charges on certain classes of ligands,‡ and also indicates the oxidation state of the central atom as the appropriate charge associated with that atom (this latter feature provides an efficient means of searching for coordination compounds as a class and searching for coordination compounds of a specific element: a metal atom with a non-H connection and an indicated charge).

---

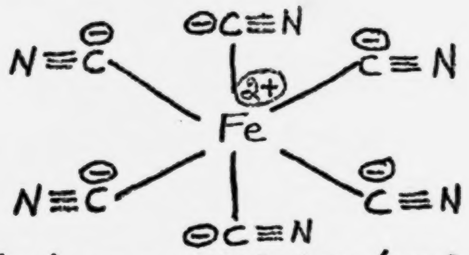
‡ See Glossary for definition.

The following are some specific coordination compound conventions:*

1. An acyclic single bond is to be used to show the connection between the metal and the group connected (ligand).

2. Stereochemistry is expressed by the usual text descriptors, +, -, syn, etc. Additional coordination compound text descriptors for the appropriate isomerism are: antiprismatic, bipyramidal, dodecahedral, planar, prismatic pyramidal, octahedral, and tetrahedral.

3. Charges are to be shown on the metal and the appropriate atom(s) of each "ionic" ligand. Example:



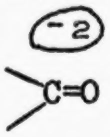
4. A hydrogen ion (H+) or a metal ion (such as Na+, Mg+2, K+) involved as a companion cation to a metal coordination anion is structured as such, in a dot-disconnected form, with the positive charge(s) indicated.

5. Coordination compounds involving charge-delocalized ligands are not presently input by structure, except for the anions of β-diketones and β-keto esters (and their thio and seleno analogs). These latter ligands usually form chelate rings with the metal

---

*For more detailed explanation and examples of structuring, see footnote on page 25.

using both oxygen (S, Se) atoms and ionizing a hydrogen atom from the carbon in between the oxygen atoms of the carbonyl groups. Therefore, a negative charge is shown on the carbon atom between the carbonyl groups, and appropriate positive charges(s) are shown on the metal.

6. The carbonyl group (CO), as a unidentate[‡] group, is represented with a carbon-oxygen triple bond ($C\equiv O$). The carbonyl group as a bridging ligand is represented as 

7. The nitrosyl group (NO) is represented with a nitrogen-oxygen triple bond and positive charge on the oxygen: $N\equiv O^{\oplus}$
8. Anions of a α-amino (hydroxy, mercapto) acids and β-amino (hydroxy, mercapto) acids usually form chelate rings with the metal (ionizing the hydrogen of the CO_2H group). The bonding is through the N (O, S) and the $-O^{\ominus}$ portion of the carboxylate group. However, such compounds involving the metals Ba, Ca, Cs, Fr, K, Li, Na, Ra, Rb, or Sr are structured by dot-disconnected conventions instead of coordination compound conventions.

Stereochemistry - Stereochemical information is recorded by text descriptors in the modification field "TEXT". The Substructure Search System retrieves all structures which satisfy the constitutional (two-dimensional) requirements of a substructure. The chemist then screens this set of potential answers for those which satisfy the stereochemical requirements as well.

[‡] See Glossary for definition.

Text descriptors have been developed for the main body of steroids, terpenes, alkaloids, and carbohydrates based upon an alphabetic term, or base name, representing a basic parent structure with implied stereochemistry at specified positions. This base name is the name of the parent structure or a term closely related to it.

2.5 Mechanized Treatment of Tautomers and Completely Conjugated Cyclics

Tautomers - The Registry System will provide programmed identification of certain types of unique compounds that can be represented by two different, but equally valid structural diagrams. A generalized representation is:

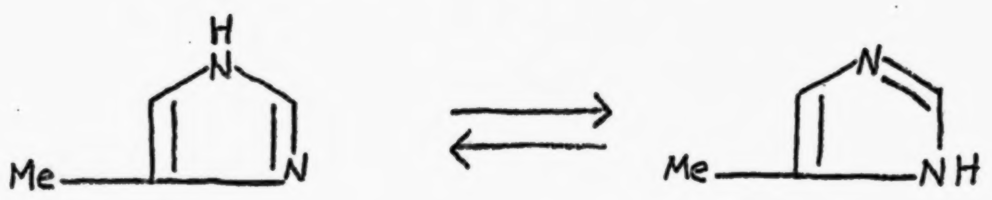
$$M=Q-ZH \rightleftharpoons HM-Q=Z$$

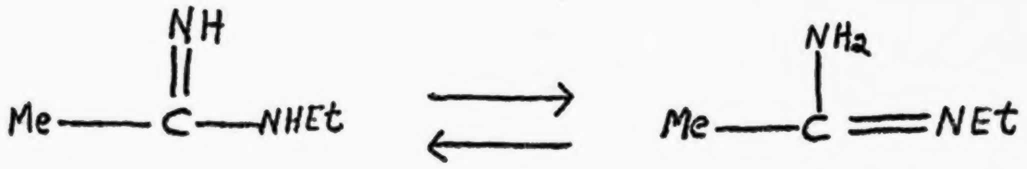
where M, Q, and Z are combinations of As, Br, C, Cl, I, N, O, P, S, Sb, Se, and Te (including abnormal mass analogs), "$=$" represents a double bond, "$-$" represents a single bond, and H must be present as shown. The following are examples of the types of structures involved:

a) General representation $M=Q-ZH$

M and Z are N (trivalent)

Q is C

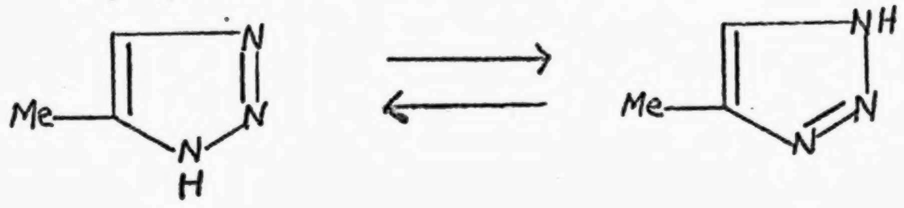


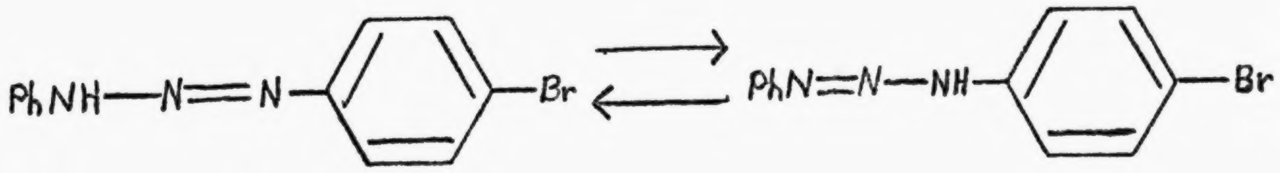


$$Me-\overset{NH}{\overset{\|}{C}}-NHEt \rightleftharpoons Me-\overset{NH_2}{\overset{|}{C}}=NEt$$

b) General expression M=Q—ZH

M, Q, and Z are N (trivalent)

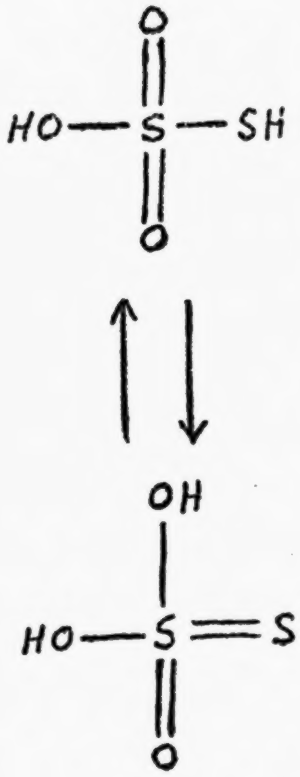


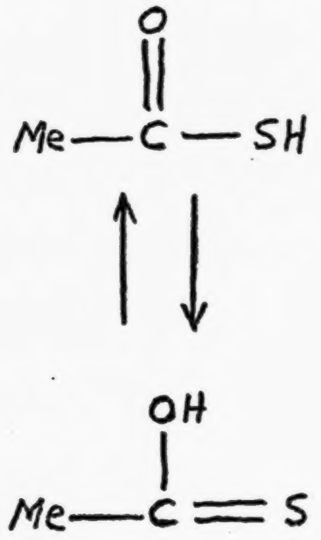


c) General expression M=Q—ZH

M and Z are O, S, Se, or Te (or abnormal mass analogs)

Q is N, P, As, or Sb (tri- or pentavalent) S, or Se (tetra- or hexavalent), Cl, Br, or I (any valence), or C (tetravalent)





These general expressions cover those forms which differ from each other in the location of a mobile hydrogen atom on the endpoints of a three atom string. The bonds involved in the string must both be either cyclic or acyclic.

When a tautomer string is recognized by program, the bonds in the string are assigned a bond value of 4 together with the appropriate bond operator. The hydrogen atom being shared by the endpoints of the string is not recorded in the H-COUNT field of the graph proper, but the endpoints of the string are recorded in a modification field designated TAUTOMER ENDPOINTS. It is understood that the two atoms in each of these tautomer endpoint pairs are sharing one hydrogen atom. Deuterium and tritium are not treated as hydrogen atoms for the purposes of tautomerism and must be accounted for in searching when they are acceptable substitutes for normal Hydrogen.

Completely Conjugated Cyclics - The Registry System accepts input of alternating double and single bonds in cyclic structures. That is, no specific indication of aromaticity, other than conventional bonding, is made at input. In the registration process, programs identify those bonds which are part of a completely conjugated cyclic system and equalize them (i.e., all bonds in such a system are given the same bond value). Such bonds are represented by the descriptor "*5" wherever they appear in the stored structure record. Whenever a completely conjugated cyclic bond so identified by program is also in a tautomer string recognized by program, that bond will be represented in all stored records by the descriptor "*5".

APPENDIX B

The Chemical Compound Registry

by

Margaret K. Park

Presented at SLA/ACS Meeting at CAS, May 4, 1968

The Chemical Compound Registry*

by

Margaret K. Park

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio

Why a Computer-Based Compound Handling System

The unifying theme of the chemical literature has been the emphasis on detailed understanding of the structural characteristics of chemical compounds and materials. The primary literature of many natural and physical sciences often has a relatively short, useful life, but even today it is not uncommon for a working chemist to go back 60 years or more in the literature to find immediately useful recorded data. Significantly, that portion of the chemical literature that has a long useful life includes data associated with substances having identified structure and/or composition. The indexes of the secondary publication in chemistry make access to the compound data readily available. In this respect, the 3 to 4 million compounds already recorded in the past volumes of Chemical Abstracts and Beilsteins Handbuch der Organische Chemie represent a valuable collection of information to the chemical community.

---

* Presented at a Joint Meeting of the Special Libraries Association Chemistry Division and the American Chemical Society Chemical Literature Division. May 4, 1968.

NOT FOR PUBLICATION

Knowledge of molecular structure provides a wide range of information about chemical reaction, physical properties, and biological activity. And, with understanding of structural chemistry rapidly growing in importance, chemical problems are increasingly being approached from the structural viewpoint. However, at the present time, many potentially valuable correlations between structural features and chemical and physical properties are not attempted because it would require too much human effort to select and compare the required features in large data collections. Moreover, existing systems in the traditional form of books or printed files cannot be easily reoriented to meet new needs. To combat these problems, computer-based methods are being called into play. The versatility of the computer in organizing and correlating the data makes possible greater flexibility in use of the files, and thus ensures a long useful life for the stored information.

Background on the Choice of a Notation

A prime concern of Chemical Abstracts Service in developing its computer-based compound handling systems is to design a system sufficiently complex in handling detailed data to provide good service to the chemical community yet simple enough in operation to be economically feasible.

The extremely large size of the operation cannot be overemphasized as an important factor in the design of the system. The Registry System has been in operation since 1965 and now has computer files of over 900,000 different chemical compounds and their associated one million names and 1.7 million bibliographic citations. These 900,000 different compounds represent the processing of 1.75 million references to compounds occurring in the literature. The Registry System has identified about half of the scattered references as being duplicates of compounds already on file. Some 4,000 new compounds are added to the files each week; reference citations to the compounds appearing in the current literature are being added at the rate of 9,000 per week.

Such a sizeable collection must be compatible with other collections of compound-related data, regardless of their form, if it is to be useful. For this reason, CAS has, as a service to the Chemical community, designed a system large enough to handle the full range of chemical literature, detailed enough in its own computer records to provide for multiple uses of the information and flexible enough to accommodate the wide variety of users who depend on these services.

Purpose and Organization of the Registry System

The Registry System is a computer-based identification system which uniquely identifies chemical compounds on the basis of their structural diagrams. The heart of this system is a computer program which generates a unique notation for each different compound recorded in the Registry files. Each machine notation is a detailed description of the two-dimensional graph of a chemical compound in an atom-by-atom, bond-by-bond listing. This chemical structure notation record is consistent with the amount of detail used by most working chemists. Further, it can be adapted to contain more detail, such as bond length or bond angles, should this be desirable, and it can also be adapted to provide lesser detail for interfacing with systems that do not use atom-by-atom records.

There are at the present time three principal files of data contained in the Chemical Compound Registry System--the atom-and-bond records just described in the structure file; a file of various forms of nomenclature that have been associated with these compounds; and bibliographic citations to the CAS publications such as Chemical Abstracts (CA) and Chemical-Biological Activities (CBAC). Registration of compounds from CA began on a routine basis with CA volume 62 (January-June 1965) and has continued since that time. In addition, the fluorine-containing organic compounds indexed in all volumes of CA from volume 1 to the present and in Beilsteins Handbuch der Organische Chemie have been registered. Also included in these files are compound data from many well known reference works such as the Merck Index, the Colour Index, some sections of the Code of Federal Regulations, United States Adopted Names, and several other references. The information in the three Registry files--and indeed in all other CAS computer files that contain compound-related data--is tied together by Registry Numbers. A Registry Number is assigned to each structure on the basis of its unique notation when the structure is first entered into the file. Whenever a structure which is already present on the structure file appears in a new source, the previously assigned number is recovered automatically. The Registry Number functions as the machine address within the associated files of structure, nomenclature, bibliographic data, index entries, etc. The number is not itself a notation that can be translated to the structural record.

Initially, the system was designed to handle carbon-containing organic compounds for which a fully defined two-dimensional structure could be drawn. Since that time, research and development work on both the chemical aspects of the Registry System and on the computer equipment and programming aspects has made it possible to extend the registration process to mixtures, inorganic compounds, metal coordination compounds, and to some classes of polymers and partially described structures.

Computer Requirements for Notation

Because each compound filed in the CAS Registry System is assigned its own unique machine-readable notation, it is appropriate at this point to review the requirements for a structure notation. Webster defines a notation as "a system of characters, symbols, or abbreviated expressions used in an art or science to express technical facts, quantities, or other data". In an extensive information system, the processes of identification, filing, and retrieving compound-related data should be handled automatically by computer processes as much as possible. It is particularly important that the CAS system, which may eventually include several million chemical compounds in one file, identify synonymy between alternate representations of compounds. The ultimate size of the file alone dictates the desirability of eliminating unnecessary, redundant information. But more important, the efficient and economic retrieval of information requires the ability to identify synonymy, whether it is for the purpose of collecting all data at a single point in listings such as the CA Subject Indexes or for retrieving data via a computer search. Effective identification of synonymy requires a unique representation for each compound in the system.

The notation for CAS use must also be a stable representation. Because it is to be used in a large, production-oriented system, the rules governing the notation must be under the control of the operating organization to assure the integrity of the data files. An evolutionary development, such as occurs with nomenclature, causes continual updating of the record as new rules are applied. For large files it is not economically feasible to continually update the collection.

The system must also be highly reliable--it must produce the same notation for the same compound no matter what the variables in input. It must be "fail safe" against keyboarding errors and inconsistencies in input. The notation should be concise and not redundant, consistent with the level of detail stored. At the same time, it should be comprehensive enough to cover all of chemistry and all classes of compounds. Finally, the notation must be flexible enough to interface with the various techniques used throughout the chemical community; useful in a variety of applications; and economical to generate, store, and utilize in the large scale operational environment.

CAS Machine Notation

The CAS notation is a detailed inventory of the atom and bond components of the structural diagram which are computer-arranged into a unique cipher.

This inventory, usually called a "connection table," is an ordered listing of the atoms and bonds, and of the drawing and the manner in which they are connected.

The computer program that generates the unique form of the cipher is the very heart of this system. This program was developed by Morgan of CAS based on the work of Gluck at DuPont. It has been mathematically verified to assure that it does, in fact, always result in the same notation for a given compound. Moreover, this technique also handled any graph of points connected with lines; therefore, all chemical compounds discovered in the future can be converted to machine notations with no modifications of the existing rules.

Stereochemistry

Third-dimensional features are recorded in Registry files by conventional stereochemical descriptors such as erythro and threo, D and L. These features are determined from the original document by the chemist who prepares the structural diagram. Methods for recording node-by-node stereoisomerism within connection table itself have been described by Petrarca, Rush, and Lynch.[1] These techniques are being evaluated to determine their economic feasibility in relation to the number of compounds in the literature that provide stereochemical detail sufficient to warrant their use.

Methods of Input-DATA

Alternative methods of input to the structure file help to build a flexible registration and filing system. For greatest efficiency, the system must input data in many forms and be adaptable to the use of many types of equipment. All processing programs in the Registry System, therefore, are intentionally designed to operate independently of the input description of the compound's structure and of the equipment which processes the data into machine readable form.

Connection Table

The structural records can be input using a variety of data records. Structural diagrams, such as those typed on structure typewriters, are input to the Registry System. Atom-bond connection tables can be entered directly into the system. Algorithms for translating systematic nomenclature and other forms of notations have been described in the literature. Regardless of which form is used, the result is an atom-bond listing of the structural diagram.

Nomenclature Translation

Computer translation of systematic nomenclature to the connection table record should provide an economical method for adding to the computer files the three to four million compounds that appeared in the indexes prior to 1965. Up to now, almost all of the compounds in the Registry files have been input via structural diagrams which were hand drawn by professional chemists and translated into machine language through clerical effort. The task of registering the pre-1965 material, however, will not be feasible--in terms of time, manpower, or dollars--if it is to require the preparation of a structural diagram for each compound and the subsequent conversion of that structural diagram to machine language for registration. Therefore, CAS Research and Development staff has concentrated on the development of translation procedures which permit direct computer translation from nomenclature to connection tables for use in automatic registration. This work--called Nomenclature Translation--has resulted in an algorithm or set of procedures that will allow the handling of an estimated 60% to 70% of the names of organic compounds in current and past CA Formula Indexes. Programming is now well underway at CAS.

Methods of Input - Equipment

A wide variety of equipment can be used to convert structural data to machine-readable form. Keyboarding can be done on devices such as the standard keypunch, paper-tape-generating typewriter, and on various magnetic type recording devices. Structure typewriters, such as those developed by Mullen[2] and Feldman[3], are also in use. A modified version of the Mullen typewriter is presently being used in CAS operations for much of the routine input to the Registry. These devices have all been used within CAS operations, often simultaneously.

Optical scanning equipment has been developed by Badische Anilin- und Soda-Fabrik to automatically scan structures into the computer and convert them to connection tables. CAS has explored optical scanning, but the equipment required to support our high-volume input is not available.

The form of input chosen and the equipment used is determined by the use to which the data is to be put. The large-scale input required at CAS dictates that the method chosen must be quick, efficient, and precise. Such a concentrated input of compounds from the full range of chemical literature justifies special purpose equipment, such as structure-typing typewriters, which are not needed for smaller operations. However, CAS will always maintain the flexibility to accommodate a wide range of equipment and input forms.

Automatic Editing

To maintain a store of accurate information in the Registry file, a computer editing program has been developed which automatically detects errors introduced during the structuring and keyboarding operations. The detailed connection table allows a relatively simple and straight-forward approach to editing this notation. Two important characteristics of the record are used for the programmed edit checks: (1) the high degree of redundancy CAS has imposed on the input conventions, and (2) the syntax of the chemistry inherent in the notation.

Using these tools, the editing program applies some 50 different checks to the connection table records. Each atom is checked to verify that it is a valid atom element symbol and that the value of the bonds attached is equal to a stored value for that element. "Uncommon" element valences, such as those that occur in free radicals and carbenes, are always recorded in the input data. Additional editing checks have been added to the 360 program to check the valence and charge of metal ions as well as the validity of the coordination number (i.e., the number of ligand attachments). The atomic mass of isotopically labeled elements is compared with a table of permissible values. Similar validity checks have also been incorporated for frequently occurring stereochemical descriptors.

From the information given in the table, the program also computes a molecular formula, including the hydrogen count, and compares it to that calculated by the chemist and input by the keyboard operator. Any discrepancy between the two molecular formulas constitutes an error. When any error is detected by the edit checks, the table is barred from further Registry processing, and the table is rejected with an appropriate diagnostic message. After being corrected the connection table is again recycled through the edit program to assure that no new errors have been introduced.

The editing program also includes features that detect alternate structural representations for certain classes of compounds and convert the different forms to consistent, unambiguous descriptions for subsequent generation of the unique notation. One such class of compounds is the conjugated ring system, such as the resonance structures of the dichlorobenzene derivatives. The alternating single and double bonds enclosed in any cyclic system are identified, and the bonds recoded as normalized, equivalent bonds. Automatic identification of such conjugated ring systems equates the alternative input records and provides an unambiguous description for the compound in the computer record yet allows the chemist to draw the diagram in the conventional manner. In a similar manner, tautomeric systems are also automatically recognized and recoded. Harmonization (normalization) of tautomeric compounds has been deliberately limited to types of tautomeric structural units found in such compound classes as the amidines, thio carboxylic, phosphoric and arsonic acids, guanidines, and imidazoles. An illustration of these types is the benzimidazole structures on the slide. Keto-enol tautomers are not equalized in this manner, although the computer routine that identifies these structural conditions is a general one that can be extended to handle this and other tautomeric classes if desired.

Another feature of the editing program adds greater specificity to the input record of the structure. This program identifies and differentiates acyclic and cyclic bonds and assigns the appropriate code to each. This additional detail provides improved discrimination in the retrieval of compounds on the basis of structural characteristics.

The Notation File

A master file of the unique notations is maintained as a part of the CAS Registry. As each new notation is added to the master file, the next available Registry Number is assigned to that notation. All subsequent entries of the same compound result, of course, in an identical notation that is compared with the master structure file to show that it is the same compound. When the notations match, the previously assigned Registry Number is retrieved.

The form of notation stored on the master file is more compact than is the connection table used for input. Redundant information, which is useful for editing the keyboard records, is not necessary after the table has been verified and accepted for generation of the unique notation. The redundancy then is removed before the notation is stored in the master file. The organization of the file itself permits an even more compact record. The lexicographic ordering of the file brings together groups of compounds with similar structural features. All acyclic structures, for example, are placed before any cyclic structures. Further, all compounds are grouped together which have the same graph. Thus, the hierarchy of notation order permits elimination of the duplicate portions of adjacent records to decrease storage space requirements. Compaction of this type decreases the file to approximately 64% of the size it would be if the entire notation were stored. In terms of file size, more than 100,000 compounds can be stored on each reel of magnetic tape. Thus, a single reel of tape will suffice for small to medium size files.

Automatically Generated Cross References

One significant advantage in a file organized into groups of closely related structures is the ability to automatically generate cross references. All registered stereoisomers occur adjacent to the non-stereospecific structure on the file. Isotopically labeled compounds are grouped with the unlabeled isomer. Amine salts, like the hydrohalides, are cited under the structure of the amine, while all metallic salts of an organic acid are grouped with the free acid. File hierarchy and the organization which it permits have helped in developing the cross reference feature of the faceted numbers which appear in Chemical-Biological Activities. Within any series of related compounds the faceted numbers for each isomer or salt include the Registry Number of the patent. Faceted numbers do not replace the Registry Numbers, but serve as cross references between acids and their salts, bases and their salts, groups of quaternary compounds that have a common cation, and stereochemical isomers.

Applications for Substructure Search

The Registry structure file provides not only the means of uniquely identifying structures, but also the data base for substructure search. The notation record permits the direct application of many substructure searching techniques. Since the notation explicitly includes the elements, their connections and the bond values and bond types, no expansion of the notation is required to effect an atom-by-atom search record. The Substructure Search System locates within the structure file or a subfile all compounds that share particular substructures or structural fragments.

The basic search system has been designed for maximum flexibility, in both the degree of specificity of the questions which can be posed to the system and the degree of specificity of the answer desired by the questioner. For example, one questioner may desire exact answers to his substructure search, while another user may want a set of answers which includes exact answers plus a selected number of closely related answers. This second option allows "browsability" which is a creative stimulus to the researcher.

Of course, economics, too, play a part in determining the most desirable level of results.

For greater flexibility, the search system is designed to permit several levels of retrieval specificity. A fragmentation search technique analogous to the widely used manual and punched card systems provides an economical retrieval tool. Chemical fragment screens, many of which correspond to the traditional functional groups, are generated automatically by computer program from the structure notation record. These types of structural screens are shown on the slide, where the traditional chemical functional groups include such groups as carbonyl, nitrile, amide, sulfonic acids, and derivatives. Some generic structural features are included for such classes as hydrocarbons, halogens, and metals.

The screening of the structural records via the fragmentation search is a very rapid, essentially tape speed search. The search program operates on all Boolean logic parameters (and/or/not) which permits considerable flexibility in search strategy. All compounds obtained as answers satisfy the search parameters, and under no circumstances are answers excluded. That is, there is 100 percent recall by this search technique. As in any fragmentation code, however, the interconnections between the structural fragments may not match the requested substructure, so that the set of answers probably includes not only all exact answers but also some closely related compounds. The degree of relevance depends to a large extent on the nature and specificity of the substructure query. Experience with files of 20,000 to 55,000 compounds indicates a 75 to 80 percent relevancy for most types of questions, which is usually quite satisfactory for most users of this high speed, relatively economical search. However, the number of answers obtained determines in many cases whether or not this degree of relevance is satisfactory--80 percent of 10 answers means that only 2 of the retrieved compounds are irrelevant, but 80 percent of 1000 or 10,000 answers is quite another matter.

For searches that do require 100 percent relevance as well as 100 percent recall, a second search technique is available that compares each atom and bond of the substructure with the connection table of the structures on the master Registry File. This iterative atom-by-atom matching technique can be used independently of fragmentation search, but it is more economical to use the rapid screening technique first to select compounds that are potential answers. The specific answers can then be identified in this subset by the slower but exacting iterative matching process. The hierarchy of the structure file organization also facilitates the retrieval since closely related compounds that satisfy the search request can be retrieved without interrogating each individual notation.

Questions may be posed to the system for either the Fragmentation or atom-by-atom search levels in terms of "and," "or," and "not" logic. "And" logic requires the presence of an atom or group of atoms in the answer. "Not" logic specifies that an atom or group of atoms must not be present in the answer. "Or" presents alternatives which may occur within a substructure or alternative substructural units that may occur in the answers. The fourth possibility, "don't care", allows atoms and bonds within the substructure to be left unspecified. The same logic operators can also be applied to two or more substructures within the same query. For example, it can be specified that two substructures must not co-occur in the same molecule.

Uses of the Substructure Search System

The potential uses of this substructure search system are quite varied.

1. It provides the mechanism for automatically generating fragmentation codes for any one of the several fragmentation search systems presently in use. Such a list of fragments can also be considered a profile to be used in automatically updating a user's structure file or fragmentation file. The results of the automatically generated fragments can be retained for computer searches or can be printed as index-type listings for manual searches. This feature has been used to assign structural class terms, or MeSH terms, used in the National Library of Medicine's MEDLARS System, where the search results are the nomenclature terms corresponding to structural fragments of the compound. Highly specialized or customized fragmentation codes can be automatically added to the high speed screen level search record by running a one-time iterative search for the structural fragments. The desired fragments are simply coded as routine search requests and the answers permanently recorded as a fragment for use in subsequent searches. And changing the fragmentation codes is as simple as redefining the substructure search and performing a single search.

Substructure Search also provides the ability to select from a large file of compounds a smaller set containing specific fragments. Thus, all compounds having significant structural features pertinent to an organization's research interest could be identified and selected as a subfile. Such a list of fragments can also be considered a profile to be used in automatically updating a user's structure file or fragmentation file.

2. Since new compounds can be identified during the registration process, this system can provide an alerting service for new compounds containing specified substructures of interest to a user. New molecular ring systems imbedded within compounds have been identified in this manner to identify ring systems for supplements to The Ring Index.

3. The search system provides customized searches for compounds which may have similar physical or biological properties because they have similar structural characteristics. Registry Numbers for compounds identified by Substructure Search can be used as parameters for the computer-based text searching in conjunction with conceptual terms and authors' names in the computer tape or printed versions of CBAC and POST.

Custom searches are by no means limited to the CAS computer--they can be conducted by other institutions or organizations using their own equipment and files, and probably in conjunction with other internal data files.

4. The search system provides the means for identifying and retrieving generic classes of compounds which can be combined with text material in handbooks or indexes for use as desk references. CAS is now developing

a computer-driven photocomposition system that will rapidly format intermized text material and structural diagrams with high graphic arts quality.

Applications of the Registry System

The information contained in the Registry files can be put to many uses. For example, the registration process itself identifies new compounds as it assigns the Registry Number, that is, it establishes the fact that the compound has been previously registered and therefore already has a Registry Number. The full benefit of the system for new compound alerting can not be derived until the records include all chemical compounds that have been reported in the literature. Only with a complete record, for example, will it be possible to determine with certainty that a given compound is or is not new to chemistry. However, the present files do provide a valuable tool for determining whether or not a compound has appeared in the literature since 1965, with manual searches of the CA indexes used in the conventional manner to scan the earlier literature.

Nomenclature

The Nomenclature File extends the uses of the familiar printed CA indexes. This File includes systematic names, as illustrated by the *Chemical Abstracts* index name, as well as trade names, generic names, and established laboratory numbers. Each name is linked to the appropriate structure by the Registry Number. These synonyms for a compound provide a greatly expanded thesaurus of terms for use in searching such publications as CA, CT, CBAC and POST. Indexes have been produced directly from the computer files through the use of computer-programmed rules for alphabetizing the compound nomenclature. A combination of names with molecular formulas provides molecular formula indexes. Another example of a molecular formula index prepared from the CAS files is an index produced for the NLM. The order of the element symbols within each molecular formula is the NOPS--nitrogen, oxygen, phosphorus, sulfur, followed by other elements in alphabetical order--rather than the modified Hill form used in CA. This NOPS sequence is generated by computer programmed instructions from the Hill form which is used within the CAS files; thus, no professional or clerical time is necessary to accomplish the translation on a routine basis.

The versatility of the computer in reorganizing and reformating the stored information in a single form is quite apparent here. Formerly, the production of multiple indexes in different sequences required the preparation of separate card files of data, manually sequenced into the different orders. Now this task can be performed easily and quickly by computer from a single input record.

The Nomenclature File is also used as a means of entry into the system. For compounds which appear frequently in the literature, it is often convenient to add new data to the files by matching names against the Nomenclature File. For small collections of names, this matching process can be done manually from alphabetically arranged indexes. For processing large collections of names, CAS has developed computer programs for matching names, retrieving the Registry Numbers and CA index names, and bibliographic references to the files. Names which are ambiguous--that is, names identified with two or more different compounds--are flagged within the files. A match against one of these ambiguous names provides multiple retrieval for review by the chemist.

Rules have also been programmed for editing nomenclature, although these editing checks are by no means as complete as are those for the structure connection tables. The present program checks primarily for consistency of format, capitalization, and italicizations, and automatically corrects many of the errors of this type that are identified. The nomenclature filing system also provides diagnostic messages for potential problems to be solved by nomenclature specialists. Work is continuing on extending the nomenclature editing features of the computer system. The nomenclature translation programs themselves provide a powerful editing function since names which can not be successfully translated into connection tables (i.e., contain errors) are rejected by the translation program with appropriate diagnostic messages for review. One of the first applications of the Nomenclature Translation programs this year will be the generation of the connection tables from the names already stored in the Nomenclature File and comparison of the generated table with the connection table already on the structure file for that compound.

The third file of data contained in the Registry System is the file of bibliographic citations. One of the more obvious uses of this information is a bibliography for any specific compound or for groups of compounds such as would be identified through the Substructure Search System. This type of retrieval was used to obtain the references to the substructure search answers obtained in the demonstration of that system at the ACS meeting in New York in September of 1966. A less obvious use of this file is the retrieval of Registry Numbers, then nomenclature, based on a given reference--for example, the reference to a book such as The Merck Index. Selection criteria based on some 30 reference works have been used to organize data into various index formats for the Food and Drug Administration and the National Library of Medicine. Another use of the bibliographic data--one which is not yet fully operational--is the selection of compound-related data for publication of indexes in the primary journals.

Registry Numbers have been appearing in the Journal of Organic Chemistry since March 1967. The publication of the Registry Numbers in the primary publication is the first step in proving molecular formula and nomenclature indexes to the compounds within the primary publication. The link between the two systems--the parent ACS organization and its secondary publishing arm, CAS, is the bibliographic citation to the original article and the corresponding citation to the CA abstract which is the citation in the Registry files. Much of the work necessary to provide all the necessary links on a time scale consistent with the publication of the article in the journal has already been completed; the conversion should be possible when CAS has fully computerized its handling of the bibliographic citation data that appears in Chemical Abstracts.

Summary

This paper presents, in brief, an overview of the CAS Compound Registry System, its component files, and the many uses to which this new computer-based facility can be put. Perhaps the most important aspect of the Chemical Compound Registry is the unified nature of the data bank. Each of the different data elements has been identified and flagged; this permits the selection and reorganization of any of the material into a vast number of specially designed formats--either in printed or machine-readable forms, or both. In addition, any given subset can include data from other portions of the total CAS data bank. Pilot services which utilize these features are already available and in use. The full potential, however, will be realized only as imaginative chemists and chemical engineers begin to identify new and better information needs within their own programs.

Acknowledgement

The contributions of many people to the development of the CAS system is gratefully acknowledged. The work reported in literature has been used in many instances as the building blocks for this notation and the CAS compound handling system as a whole. Cooperative arrangements with several organizations continue to provide valuable assistance. The support of the National Science Foundation, the National Institutes of Health, the Department of Defense, the National Library of Medicine, and the Food and Drug Administration in much of this work is also acknowledged.

LITERATURE CITED

1. (a) Petrarca, A. E., Lynch, M. F., and Rush, J. E., "A Method for Generating Unique Structural Representatives of Stereoisomers," J. Chem. Doc. 7 (1967), 154

(b) Petrarca, A. E., and Rush, J. E., "Methods for Computer Generation of Unique Configurational Descriptors for Stereoisomers Square-Planar and Octahedral Complexes," (in press).

2. Mullen, J. M., "Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures," J. Chem. Doc. 7 (1967), 88

3. Feldman, A., Holland, D. B., Jacobus, D. P., "The Automatic Encoding of Chemical Structures," J. Chem. Doc. 3 (1963) 187

APPENDIX C

A DESCRIPTION OF THE CAS CHEMICAL REGISTRY SYSTEM

A. Introduction

B. Data Flow for the Registry System

C. Registry Form

D. Registry Numbers

Extracted from the CAS "Registry Division Operations Manual" Copyright 1968 by the American Chemical Society.

Registry System Operations

A. Introduction

Since 1958, CAS has been working toward establishing a computer-based system for handling chemical information, such as physical properties, chemical reactivities, biochemical activities, and applications.

Because of the importance of chemical structures and the need to interrelate structures and corresponding chemical, physical, and biological data, a subsystem called the CAS Chemical Compound Registry System is the first step in the operation of an over-all computer-based service.

The Chemical Compound Registry System is a machine (computer) record of chemical structures, and chemical constitution expressed as molecular formulas, names, and literature references. The process of registration includes the assignment of a unique number to each different chemical structure. This number is called the Registry Number. The Registry Number is a computer-generated nine-digit number, which is not an "intelligent" number. That is, the number does not convey any information about the structure with which it is associated. The units position is a "check digit" generated by the computer. This is a safety factor to reduce errors from miscopied numbers; it is a means for allowing the computer to reject a Registry Number which is wrong. The Registry Number is the link between the structure of the compound and all other information about the compound in the additional records which will be established in the future, concerning physical and chemical characteristics, biological and medical properties, and practical industrial uses. It is intended that eventually the computer system will have records of every chemical substance and all the useful published material bearing on each substance.

Registry System Operations

Since it is intended that the machine record will eventually include all compounds, it is important that registration be as specific as possible; that is, each structure must have a separate, different, Registry Number. This idea is followed through in that an acid and each of its salts have separate Registry Numbers - acetic acid, sodium acetate, aluminum acetate, for example, all have different Registry Numbers. Similarly, with bases - aniline, aniline hydrochloride, and aniline hydrobromide have different Registry Numbers. Optically active compounds and racemic mixtures such as *d*-mandelic acid, *l*-mandelic acid, and *dl*-mandelic acid each have unique Registry Numbers; so also does mandelic acid with no stereochemistry specified. Each specified stereoisomer of a given set, such as the 1-hydroxy-2-chlorodecalins, also receives a separate Registry Number. Labeled compounds, radioactively or otherwise, receive Registry Numbers different from those of the normal compounds, such as toluene, toluene labeled with carbon-14 in the meta position and toluene labeled with tritium in the para position. Biochemical literature often furnishes information about an ion, such as lactate, acetate, or pyruvate, and in such a case the ion receives a Registry Number different from that of the parent acid or a specific salt. Geometrical isomers such as *cis*-stilbene and *trans*-stilbene have different Registry Numbers; so also does stilbene with no stereochemistry specified.

Registry System Operations

B. Data Flow for the Registry System

Registry operational flow is illustrated in a simplified diagram as shown in Figure 1. (Numbered points are described below.)

The structures which are recorded by computer processes originate as shown in Figure 1 (top center) (1) in the preparation of structures for compounds selected for index entries in the CAS Subject and Formula Indexes and for compounds appearing in digests in the journals Chemical-Biological Activities (CBAC) and Polymer Science and Technology - Journal and Patent Sections (POST-J and POST-P); (2) from other sources, CAS files, non-CAS files, reference works, etc. -- in some cases the structures are already present (as in the Merck Index) and in other cases the structures must be supplied, usually by the Formula Indexing Department.

It should be noted that, in the routine work flow involved in CA Indexing and Registry Divisions, the structures prepared in the Formula Indexing Department pass into the Registry Division for the purpose of registration and then proceed to the Subject Indexing Department so that the indexer can supply an index name for the compounds.

1. Information enters the Registry System on specially designed Registry Forms, one for each compound to be registered. As it leaves the "structure drawing operation", the form contains a preprinted temporary identification (TID) number and the following information supplied by the originating chemist: the structural formula to be registered, the molecular formula (often abbreviated "molform"), the nomenclature*

---

* Nomenclature includes systematic and nonsystematic names, acronyms, laboratory numbers, etc.

(not routinely, but in some cases), the bibliographic citations to information about the compound, and a term or descriptor that describes the stereochemical, labeling, and, in some cases, other aspects of the structure. (This descriptor, presented later in detail, is one of the conventional stereochemical descriptors used in the chemical literature, or a device to differentiate further between structures.)

2. Structures are checked and classified by the originating chemists as to the necessary mode of registration.

3. The Registry Forms are sorted and batched. Most compounds are machine registered, but manual registration (cf. point 15) is used for: (1) compounds with more than 255 non-hydrogen atoms, and (2) compounds for which structuring conventions have not yet been established for the Registry System.

4. The connection tables or structures, molforms, source codes, and text descriptors are keyboarded and processed according to point 7 below.

5. The Registry Forms are placed in the Inwork File.

6. Registry Forms remain in the Inwork File until registration is complete for the compound. They are then added to the Master File of Registry Forms (cf. points 9 and 10).

7. The information keyboarded in point 4 above is entered into the computer, where the compound is either (1) registered (cf. point 8); (2) added to a listing of compounds that have the same "two-dimensional"

structure, but differ in textual descriptors (cf. point 11) or; (3) rejected because of keyboarding or connection-tabling errors (cf. point 13).

8. As compounds are registered, records are kept of compounds new to the system and compounds that have been registered previously. Gummed labels containing the assigned Registry Number are printed for all compounds. Previously registered compounds are identified by a series of asterisks which precede the TID on the label.

9. The gummed labels are placed on the proper Registry Forms in the Inwork File.

10. Registry Forms are filed, in Registry Number order, in the Master File of Registry Forms.

11. If a connection table for a structure being processed is the same as one or more previously registered structures, but has a different textual descriptor, then all structures in question are listed for resolution by a chemist.

12. With the aid of Registry Forms pulled from the Inwork and Manual Files the chemist resolves the problems determining which structures are identical and which differ in their stereochemistry. Resolved problems are re-entered into the system at point 7.

13. Compounds rejected for keyboarding or connection tabling errors are reviewed clerically, and errors that are detected are corrected and re-entered into the flow at point A in the diagram. Errors that cannot be resolved clerically are sent to a chemist for review.

Registry System Operations

14. The reviewing chemist either corrects the error and re-enters the Registry Form into the system at points A or B, or, for problems he cannot resolve on the basis of the information available to him on the Registry Form, returns the Form to its source (structure drawing operation) for resolution of the problem.

15. Compounds that cannot be machine registered are registered manually by a chemist. For this purpose, several files of manually registered compounds are maintained either in molform order or in name (alphabetically) order, or in Registry Number order. The chemist either retrieves from the file the Registry Number previously assigned to the compound or else assigns a new number if the compound has not been registered. The TID and Registry Number of the manually registered compounds are keypunched and added to the machine files.

16. The TID, molform, nomenclature, and bibliographic citations are input into the Data Sheet system. They are sent to data input routines for the Bibliography System.

17. Name Match Registration process - A system whereby the Nomenclature used to identify substances to be registered are keyboarded and automatically matched against the All Name File to recover previously assigned Registry Numbers and Index Names and to permit updating of the Bibliography File.

DATA FLOW FOR CHEMICAL COMPOUND REGISTRY

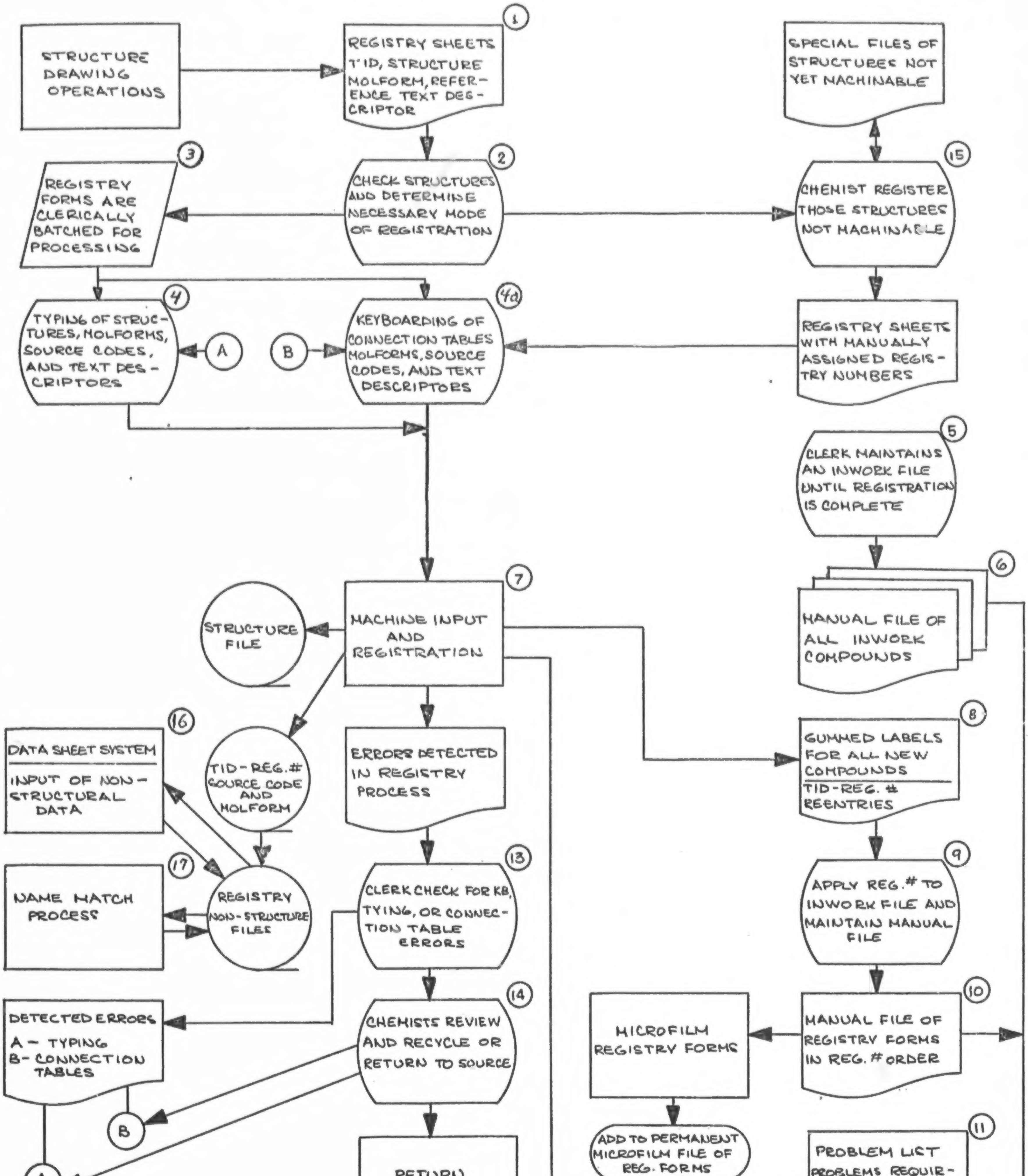

STRUCTURE DRAWING OPERATIONS
①
REGISTRY SHEETS T'ID, STRUCTURE MOLFORM, REFERENCE TEXT DESCRIPTOR
SPECIAL FILES OF STRUCTURES NOT YET MACHINABLE
③
REGISTRY FORMS ARE CLERICALLY BATCHED FOR PROCESSING
②
CHECK STRUCTURES AND DETERMINE NECESSARY MODE OF REGISTRATION
⑮
CHEMIST REGISTER THOSE STRUCTURES NOT MACHINABLE
④
TYPING OF STRUCTURES, MOLFORMS, SOURCE CODES, AND TEXT DESCRIPTORS
A
B
4a
KEYBOARDING OF CONNECTION TABLES MOLFORMS, SOURCE CODES, AND TEXT DESCRIPTORS
REGISTRY SHEETS WITH MANUALLY ASSIGNED REGISTRY NUMBERS
⑤
CLERK MAINTAINS AN INWORK FILE UNTIL REGISTRATION IS COMPLETE
⑥
MANUAL FILE OF ALL INWORK COMPOUNDS
⑦
STRUCTURE FILE
MACHINE INPUT AND REGISTRATION
⑯
DATA SHEET SYSTEM
INPUT OF NON-STRUCTURAL DATA
TID-REG. # SOURCE CODE AND MOLFORM
ERRORS DETECTED IN REGISTRY PROCESS
⑧
GUMMED LABELS FOR ALL NEW COMPOUNDS
TID-REG. # REENTRIES
⑰
NAME MATCH PROCESS
REGISTRY NON-STRUCTURE FILES
⑬
CLERK CHECK FOR KB, TYPING, OR CONNECTION TABLE ERRORS
⑨
APPLY REG. # TO INWORK FILE AND MAINTAIN MANUAL FILE
DETECTED ERRORS
A - TYPING
B - CONNECTION TABLES
⑭
CHEMISTS REVIEW AND RECYCLE OR RETURN TO SOURCE
MICROFILM REGISTRY FORMS
⑩
MANUAL FILE OF REGISTRY FORMS IN REG. # ORDER
B
A
RETURN TO SOURCE
ADD TO PERMANENT MICROFILM FILE OF REG. FORMS
⑪
PROBLEM LIST PROBLEMS REQUIRING CHEMICAL RESOLUTION
⑫
CHEMIST RESOLVE PROBLEMS, GENERATE FEEDBACK TO REGISTRY

Figure 1

Registry System Operations

C. Registry Form

A form called the Registry Form is used for the structures, a separate form for each structure. This form has undergone several revisions; the latest is shown as Figure 2. The information given on this form, if the source is the current CA volume, is for the use of the Subject Indexer as well as Registry Division personnel.

A description of the form follows:

1. ID (Identification) - Space for control information relative to generation of the sheet.

2. Ref. Source (Reference Source) - The source of the information on the sheet.

3. Vol. (Volume) - The CA volume number, or alternatively the names or abbreviations of whatever the source is for the structure.

4. Col. (Column) - The CA page (column) number and fraction where reference is made to the compound in the CA abstract, or alternatively, the page reference used in another source.

5. Abt. (Abstract) - The number assigned to the CA abstract. The abstracts in a column are numbered in sequence.

6. Compd. No. (Compound Number) - The number assigned by the Formula Indexer to the compound in the particular abstract, original paper, or unit of work. Each compound selected for indexing within a given original publication is given a number in sequence from one to the final number of compounds in the publication.

7. Sheet No. (Sheet Number) - The consecutive working record number of the chemist preparing the structure.

Registry System Operations

8. Chem. (Chemist) - (a) The person who prepares the structure and calculates the molecular formula; (b) the person who provides the CA index name for the compound.

9. Codes - The mode code (a letter) given first which indicates the input method (R = connection table; no other codes have been assigned yet), and the source code (a two-digit number) which indicates the source (for example, 16 is the code for CA Volume 63). R16 indicates that the structure is from CA 63 and has been recorded by connection table.

10. Reg. No. (Registry Number) - Assigned out of a handbook (for example, the Name Index) or as a result of mechanized recording.

11. Added Ref. File Codes (Added Reference File Codes) - Multiple references for a compound appearing in source documents following the first reference (point 2).

12. Names - Name(s) or designations (such as a Roman Numeral) of the compound as provided by the author of the original paper.

13. TID (Temporary Identification Number) - A mechanically generated number, preprinted on the Registry Form, used to identify a compound until the Registry Number is assigned.

14. TID - Space for typing the TID from 13.

15. Prop. (Properties) - Physical properties noted in the abstract or original, for the aid of the Subject Indexer.

16. TID - Identical to the TID of point 13, but appearing in the area to be used for possible optical scan input of the structure.

17. R (Ring) - If computer ring system analysis is required for indexing use, the R is crossed out.

18. Text-desc. (Textual Descriptor) - Notation of stereoisomerism, isotopic labeling, or other information to be added in an appendix to the connection table.

Registry System Operations

19. IMF (Index Molecular Formula) - The molecular formula used by the indexer; this space is not filled in when the IMF and MF (See point 20) are the same.

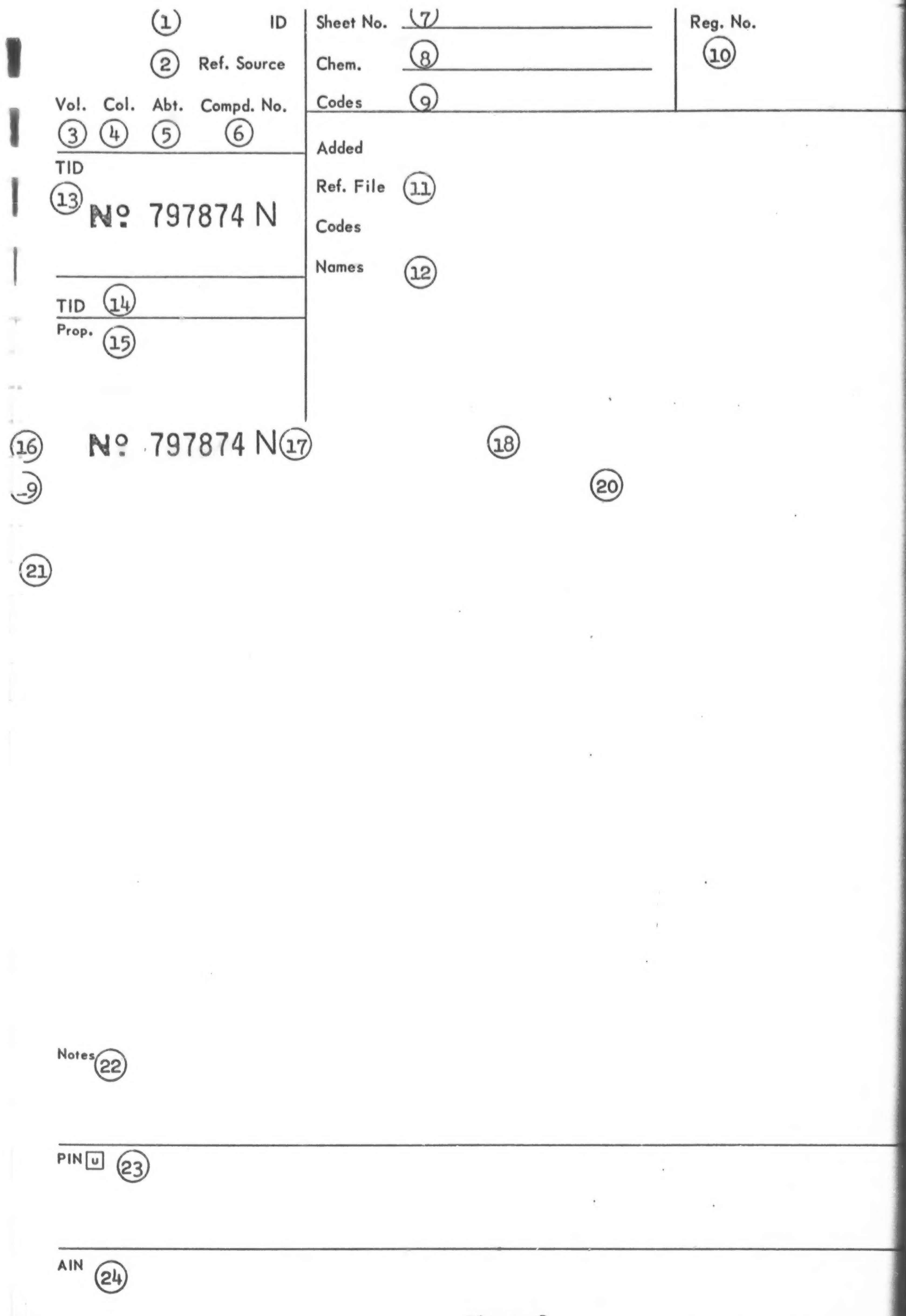
20. MF (Molecular Formula) - The molecular formula, as used in the Registry System, for the compound.

21. Structure Area (Not actually labeled - the large blank area in the center of the Registry Form) - The area in which the structure is drawn or reproduced.

22. Notes - Observations of the person who prepared the structure. They pertain to other information on the Registry Form.

23. PIN ☐U (Preferred Index Name) (Uninverted) - The preferred Index Name in the inverted form, as it appears in the CA Subject Indexes. If the name is in the uninverted form, notation is made in the U-box. This space is often vacant, since the name may not be available to the person who prepares the structure.

24. AIN (Added Index Name) - An additional CA name which appears in the CA Subject Index. For example, some complex esters are given both acid - and alcohol - based names, although the acid-based name is the preferred. This space is often vacant, since the name is often not available to the person who prepares the structure.



① ID

② Ref. Source

Vol. ③ Col. ④ Abt. ⑤ Compd. No. ⑥

TID ⑬ № 797874 N

TID ⑭

Prop. ⑮

Sheet No. ⑦

Chem. ⑧

Codes ⑨

Reg. No. ⑩

Added

Ref. File ⑪

Codes

Names ⑫

⑯ № 797874 N ⑰ ⑱

-9 ⑳

㉑

Notes ㉒

PIN ☐U ㉓

AIN ㉔

Figure 2

D. Registry Numbers

The Chemical Compound Registry Number is a nine-digit number whose units position is a check digit generated by the computer. The number has no established pattern in that no part of the number corresponds with any feature of the structure to which it is assigned.

The first eight digits of the Registry Number are a number serially assigned by computer at the time of registration. (The eight digits presently include from one to three zeros for the hundred-thousand, million, and ten-million positions.) The ninth digit (the units position) is a check digit used to reduce transcription errors. The check digit is computed on the basis of the first eight in the following manner. The position of each digit of the eight-digit number is numbered from right to left starting with the number one, without skipping. Each digit of the number is multiplied by its position number and the results are added. The digit in the units position of the final result is the check digit and becomes the units digit of the Registry Number.

For example, assume the number serially assigned is 00095216; numbering the positions gives:

87654321
00095216

Multiplying each digit by its position and adding the results yields, from right to left:

$$(6 \times 1) + (1 \times 2) + (2 \times 3) + (5 \times 4) + (9 \times 5) + (0 \times 6) + (0 \times 7) + (0 \times 8)$$

$$= 6 + 2 + 6 + 20 + 45 + 0 + 0 + 0$$

$$= 79$$

Registry System Operations

The units digit of the answer is 9 and this is the check digit of the number 00095216. Adding the check digit to the right end of the number gives 000952169, the Registry Number. The zeros in the leftmost positions are omitted for printing and publication purposes.

All of the computer programs at Chemical Abstracts Service which handle the Registry Number include the checking routine. Thus, for each data manipulation keyed to Registry Numbers, the Registry Number is validated. The chances of misidentifying one Registry Number for another are substantially reduced and much proofreading is avoided.

Two special series of Registry Numbers have been established, one for mixtures, and the other for polymers. The Mixture Registry Numbers are characterized by the prefix MX attached to a seven-digit Registry Number drawn from the eight million series; for example, MX8000008. The Polymer Registry Numbers are composed of the prefix PM attached to a seven-digit Registry Number drawn from the nine million series; for example, PM9000004. The machine validates the presence of the particular prefix involved.

In order to make the Registry Numbers easily identifiable, Chemical Abstracts Service has established a standard format. The format is (D = digit):

$$D_{1-6}\text{-DD-D}$$

From left to right, this consists of a group of one to six digits and an alphabetical prefix if present, a dash, a group of two digits, a dash, and the last digit. For example, 89-96-3; 3345-05-9; MX8264-01-9; PM9016-24-4.

APPENDIX D

The Generation of a Unique Machine Description for Chemical Structures -- A Technique Developed at Chemical Abstracts Service

by

H. L. Morgan

[Reprinted from the Journal of Chemical Documentation, 5, 107 (1965).]

Copyright 1965 by the American Chemical Society and reprinted by permission of the copyright owner.

The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service

H. L. MORGAN
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received January 15, 1965

I. INTRODUCTION

As part of the development of a computer-based chemical information system at CAS, it has been necessary to devise techniques for the registration of drawings of chemical structures. A major purpose of the CAS registration process is to determine whether a particular structure has already been stored in the system. The ability to make this determination makes it possible to utilize a computer to assign to every chemical structure a unique identifying label. This identifying label, referred to as a registry number, is the thread that ties together all information associated with a particular compound throughout the developing CAS computer system. It is because of this association, made possible by the registration process, that CAS will be able to provide multiple-file correlative searches with assurance that all information on file for a particular compound has been located.

II. THE REGISTRATION PROCESS

The registration technique that has been selected by CAS requires computer generation of an alphanumeric description for each chemical structure that is unique for that structure. The machine technique has not yet been extended to all types of structural detail, but techniques and computer programs are complete for generating the unique description for the two-dimensional projection of fully known nonpolymeric chemical structures. The third dimension is presently handled by the addition of conventional stereochemical descriptors which are supplied by the chemist who prepares the structural diagram for input to the system.[1]

In the coming months present basic machine techniques will be extended to handle partially unknown and polymeric structures. Work is also progressing toward the inclusion of the third dimension directly in the graphic record so that in time the full steric picture will be in the form of a single detailed coherent record of each structure. The initial approach, however, permits CAS to provide an operable registry system that will accommodate all compounds without awaiting the utopia of a complete set of machine techniques and computer programs that will handle all chemical substances automatically.

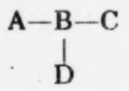
Once the unique descriptions for a set of input structures are obtained, the remainder of the registration process is simple and very fast. Since the description of each compound is in itself unique it is possible to organize both the input and registry files into a unique sequence. The use of this unique sequence reduces the actual registration process to a merging and updating of two serial files; therefore, it is the uniqueness of the machine representation of a chemical compound that is the key to an effective, efficient, reliable registration system.

III. CHARACTERISTICS OF THE STRUCTURE DESCRIPTION

The structure description employed in the CAS registration process is a uniquely ordered list of the node symbols of the structure (or graph) in which the value (atomic symbol) of each node and its attachment (bonding) to the other nodes of the total structure are described. Such a list and description is called a "connection table." Since this paper is not concerned with structure input, the connection table which is described is that stored and manipulated by the computer. The form of the table which is used within the computer is not the most convenient form for input to the system; thus the input form is translated by the computer into the "compact connection table" developed by D. J. Gluck of du Pont.[2] In this form of the table, the nonhydrogen nodes of the structure are listed according to an exact set of rules. The application of these rules alone does not produce a unique table; it does, however, produce a partial ordering among the nodes of the structure. "Partial ordering" in this context means that at certain stages in the formation of the table certain nodes will receive preference for earlier listing in the table. This is important since the generation of the unique table is based in part on a process of successive partial orderings as will be seen later in this paper.

After establishing a structure representation in the computer memory, the compact connection table is formed by first numbering the nonhydrogen nodes of the structure. This numbering proceeds from 1 using only the ordinal numbers. The numbers are assigned to the nodes of the structure according to the following rules: (1) a node is arbitrarily selected and assigned the locant, node number, 1; (2) the nodes attached to node 1 are numbered 2, 3, etc. When all the nodes directly attached to node 1 have been numbered, those which have not yet been numbered but which are attached to node 2 are numbered, and so on. This procedure is followed until all nodes have been numbered, or as in the instance of disconnected graphs such as represent ions, until the process leads to a point where not all nodes have been numbered, yet none of the unnumbered nodes is attached to a previously numbered node. Under such conditions another arbitrary choice is made among the unnumbered nodes for the next node to be numbered and the process of numbering is continued.

Example I.—Assume the structure



For this structure the following table shows the numberings that result from application of the above rules.

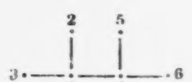
Locant	Possible node assignments											
1	A	A	B	B	B	B	B	B	C	C	D	D
2	B	B	A	A	C	C	D	D	B	B	B	B
3	C	D	C	D	A	D	A	C	A	D	A	C
4	D	C	D	C	D	A	C	A	D	A	C	A

For the structure in example I there are 24 possible numberings using the numbers 1–4; however, only 12 of the possible numberings comply with the rules cited above. This reduction is a characteristic of the numbering rules and becomes more significant as the size and complexity of the structure being treated increase.

When the entire structure has been numbered according to the preceding rules the connection table is formed by recording the structural relationships in the five lists which compose the connection table, as follows:

1. The "FROM ATTACHMENT" List.—This list is composed of X fixed length ranks where X is equal to the number of nonhydrogen nodes in the structure. In this list the ith rank is used to describe not more than one attachment between the ith node and one other node of the structure. At the ith rank is recorded the rank number of the lowest numbered atom attached to the ith node. If, however, the rank number which would be recorded at the ith rank is numerically greater than i, the ith rank is left blank.

Example II.—Assume the following structure with the numbering shown



---

(1) The system is now an operational element of the publication process of *CBAC*, a new CAS computer-based publication.

(2) D. J. Gluck, *J. Chem. Doc.*, 5, 43 (1965).

For this structure with the numbering shown the "FROM" list is shown below. The rank numbers to the left in this and following examples are for the reader's convenience and do not appear in the actual list.

Rank no.	From attachments
1	Blank
2	001
3	001
4	001
5	004
6	004

2. The "RING CLOSURE" List.—This list is composed of X fixed length ranks where X is equal to the number of cycles (rings) in the structure. Structures containing no cycles have no such list.

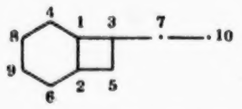
After the formation of the FROM list there will remain one connection, not described in the FROM list, for each cycle in the structure. These additional bonds or ring closures are defined in the RING CLOSURE list as follows:

(a) For each ring closure, record in a rank of the RING CLOSURE list the locants of the two atoms involved.

(b) In each rank of the RING CLOSURE list, order the two locants so that the first is numerically less than the second.

(c) Order the ranks of the RING CLOSURE list so that the locant pair of the first rank is numerically less than the second, which is less than the third, etc. Thus, 002 007 < 003 005 < 003 006.

Example III.—Assume the following structure with the numbering shown.



For this structure with the numbering shown, the FROM list and the RING CLOSURE list are as follows:

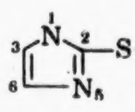
Rank no.	From attachment	Ring closure
1	Blank	
2	001	
3	001	
4	001	
5	002	
6	002	
7	003	
8	004	
9	006	
10	007	
		003 005
		008 009

It should be noted at this point that the FROM list and the RING CLOSURE list are sufficient to completely describe the interconnections of the graph for the two-dimensional projection of the compound.

3. The "NODE VALUE" List.—This list is composed of X fixed length ranks where X is equal to the number of nonhydrogen nodes of the structure. In this list the ith rank is used to describe the node value (atomic symbol) of the ith node (see example IV below).

4. The "LINE VALUE" List.—This list is composed of X fixed length ranks where X is equal to the number of bonds in the structure between two nonhydrogen nodes. In this list the ith rank is used to describe the line value or bond for the attachment defined at the ith rank of the FROM list or RING CLOSURE list. For purposes of definition, the ranks of the RING CLOSURE list are assumed to be numbered consecutively after the FROM list. The bonds (*i.e.*, line values) are described by assigned code.

Example IV.—Assume the following structure with the numbering shown.



Rank no.	From attachment	Ring closure	Node value	Line value
1	Blank		N	Blank
2	001		C	1
3	001		C	1
4	002		S	1
5	002		N	2
6	003		C	2
		005 006		1

5. The "MODIFICATIONS" List.—This list is used to describe any other modifications of the nodes and lines as listed, such as the charges of ions, isotopic mass, and citation of unusual valence.[3] Such modifications are described by citing the type of modification in coded fashion, followed by the node number or line number being modified, followed by a description of the modification in coded form. Since the techniques for treating this list are merely an extension of the techniques applied to the previous four lists, discussion of the MODIFICATIONS list will be omitted from the remainder of this paper.

The compact connection table is at this stage an unambiguous description of the two-dimensional projection of a chemical structure drawing. Where necessary it is made unambiguous for three-dimensional structures by the addition of conventional stereochemical descriptors as previously mentioned. Thus, the table is at this stage an unambiguous but non-unique machine representation of the chemical structure. It is one of a family of unambiguous descriptions of the structure. The exact table selected for use in the CAS Registry System is a member of this family and is selected by further computer processing.

In the following pages the techniques for selecting the unique table from among the family of unambiguous tables will be shown to be completely independent of the order of the nodes in the input table. Since the ordering process is independent of the order of the nodes in the input table, it follows that the unique table is also independent of both the orientation and the projection of the drawn structure. It also follows that the ordering process is independent of the means by which the drawn structure is converted to a machine representation, *e.g.*, Army Chemical Type-

---

(3) The editing routines include a check for normal valence. Thus, for example, a trisubstituted methyl free radical requires the specification that only three groups are directly bonded to the methyl carbon instead of the usual four.

writer,[4] optical scanning,[5] clerically generated connection table,[6] grid structure,[7,8] or linear notations,[9] so long as the resulting machine representation is in fact a representation of the structure in question. The points expressed in this paragraph are very important since the CAS Registry System is based on the premise that a unique structure will be stored once and only once, thus making the registry number a unique and unambiguous identification of a chemical substance.

IV. THE GENERATION OF THE UNIQUE DESCRIPTION

As has been stated, the unique table used in the CAS system is a member of a family or set of tables, all of which describe the same structure equally well. It is unimportant, therefore, which member of the set is labeled unique so long as the same table is always selected for the same structure. Since it can be shown that the set is finite for any graph composed of a finite number of nodes, it is possible to select the unique table by generating all members of the set, lexicographically ordering the members of the set based on the characters involved in the description, and then selecting the first member of the resulting list as the unique table. This concept is a restatement of a technique proposed by C. N. Mooers for generating a unique cipher based on a process of making all possible "cuts" and comparing the resulting ciphers.[10, 11]

The generation of all possible tables of the type described would, in the case of large molecules, be prohibitively expensive. It is necessary, therefore, to devise techniques to limit the number of tables that must actually be generated to some invariant subset of tables which is small enough to make the process economically feasible. Having generated this invariant subset, the unique table is selected in exactly the same way as if the entire set had been generated. It does not necessarily follow that the same table would be selected from the subset as would be selected from the entire set, but that fact is not important so long as only a single subset is generated for a given compound regardless of the order of the nodes in the input table for that compound.

In order to generate only an invariant subset of the possible set of tables, the computer program first employs the rules for numbering the structure and forming the table as described earlier. This procedure reduces what would be a factorial expression to a number which is almost always significantly smaller. For instance, in a simple six-membered ring there are 720 possible numberings; however, only 12 comply with the rules for numbering. Thus, the numbering rules have created an invariant subset. In addition to the rules of numbering, the computer program employs certain invariant properties of the graph to reduce further the size of the subset. These properties are the "connectivity value" of each node, the node value (atomic symbol), and the line value (bond).

The second means by which the subset is reduced in size is by introducting a partial ordering among the nodes of the graph. The selection of the next node to be listed, where a choice is possible, can then in many cases be resolved on the basis of a preference implicit in this partial ordering. A simple illustration of such a partial ordering is shown in example V where preference is given to the nodes with the greater number of attachments at each point of choice.

Example V.

Structure	Possibilities for the order of citation of the nodes	
A—B—C—D	B	C
	C	B
	A	D
	D	A

In example V only nodes B and C were considered for node 1 because of the preference introduced by the partial ordering. Having selected one, the other is given preference for node 2 again because of the partial ordering. Having listed nodes 1 and 2, nodes 3 and 4 are fixed because of the rules for numbering. Thus, in this example the subset generated will consist of only two tables, whereas without the use of the partial ordering six tables would have been generated, and without both the partial ordering and the rules for numbering twenty-four tables would have been required.

Although the partial ordering of the nodes based on the number of attachments will usually greatly reduce the number of tables in the subset, it is not sufficient to adequately partition the set. The reason for this is that in organic chemistry the number of bonds to any given atom rarely exceeds four or five. In order to increase the effectiveness of the partial ordering, a technique has been devised for computing a "connectivity value" for each node based on the invariant properties of the graph. These values are then used to introduce a partial ordering among the nodes in the same fashion as the number of connections were used in example V.

The "connectivity values" are computed by first assigning to each node an initial "connectivity value" equal to the number of nonhydrogen atoms attached to that node. This number is clearly an invariant property of the graph. The computer then calculates the number (k) of different "connectivity values" which had been assigned. An iterative process is then established which calculates a new "connectivity value" for each node. This new value is the sum of the assigned values for the nodes connected to the one under consideration. Having computed a new value for each node based on the previous values, the computer calculates the number (k') of different values in the set of new values. If $k' > k$, the new values are assigned to the corresponding nodes, k is set equal to k', and the summation process is repeated. If, however, $k' \leq k$ the process is terminated, and the last set of values assigned to the nodes is used to induce a partial ordering among the nodes. Using this partial ordering, the size of the subset is reduced by giving preference to the

---

(4) A. Feldman, D. B. Holland, and D. P. Jacobus, *J. Chem. Doc.*, 3, 187 (1963).

(5) W. E. Cossum, M. E. Hardenbrook, and R. N. Wolfe, *Proc. Am. Doc. Inst.*, 269 (1964).

(6) G. M. Dyson, W. E. Cossum, M. F. Lynch, and H. L. Morgan, *Inform. Storage Retrieval*, 66 (1963).

(7) P. Horowitz, and E. M. Crane, "HECSAGON: A System for Computer Storage and Retrieval of Chemical Structure," Eastman Kodak Co., Rochester 4, N. Y., 1961.

(8) W. H. Waldo, and M. DeBacker, "Proceeding of the International Conference on Scientific Information, Washington, D. C., Nov. 16-21," Washington, D. C., 1959, pp. 711-730.

(9) H. T. Bonnett, *J. Chem. Doc.*, 3, 235 (1963).

(10) C. N. Mooers, "Ciphering Structural Formulas—The Zatopleg System," Zator Technical Bulletin No. 59, Zator Co., 79 Milk St., Boston 9, Mass.

(11) C. N. Mooers, "Generation of Unique Ciphers for a Finite Network," Zator Technical Bulletin No. 49, Zator Co., 79 Milk St., Boston 9, Mass.

node associated with the higher "connectivity value" at each point of choice in the numbering process described earlier. It is important to note that the iterative process is finite for any graph of X nodes where X is a finite number. The process will terminate, under the conditions cited, after no more than $(X + 1)$ iterations since there are at most X values that can be assumed by k which will cause the process to continue. Examples 1 and 2 of Appendix I illustrate the application of this technique for introducing a partial ordering among the nodes of the graph.

After introducing the partial ordering among the nodes, the generation of the subset of tables defined by this ordering is begun. Even at this stage, however, the entire subset is not always generated since in practice it is often possible to eliminate large blocks of potential tables. To describe the means by which potential tables are eliminated during the generation process, it is necessary to describe the means by which the unique table is ultimately selected.

After generating any two of the tables of the subset, a preference between them is introduced by "alphabetizing" on the basis of the collating sequence of the machine symbols involved in the tables. The table which "sorts" to the top of the list is then selected as preferred over the other. If the two tables are identical, one is arbitrarily selected as preferred over the other. For purposes of this "alphabetization," the tables are treated as a string of symbols in the following order (see example 3 of Appendix I):

A. The "FROM ATTACHMENT" list
B. The "RING CLOSURE" list
C. The "NODE VALUE" list
D. The "LINE VALUE" list
E. The "MODIFICATION" list

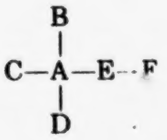
Since a preference or a lack of preference is introduced each time two tables are completed, it is never necessary to have more than two complete tables in memory at any given time.

During the table generation process, when a complete table is in the computer memory and a second table is being generated, a determination is made after each step in the generation process whether the first, completed table is already preferable to the second, partially generated one. This determination is accomplished by comparing the two FROM lists up to the point of completion of the second and selecting, as preferred, the one which "sorts" first. If it is determined that the completed table is already preferred to the second, further generation of the second table is stopped, and all tables based on the fragment thus far generated are eliminated.

Another means of eliminating potential tables during the generation process is the provision of performing a simple look-ahead to determine a preference or lack of preference.

In chemical structure drawings it is quite common to have two or more terminal atoms attached to the same atom. (A terminal atom is here defined as an atom attached to only one nonhydrogen atom.) The partial ordering of the nodes as described above does not resolve the order of selection of these terminal atoms. Thus, without provision for a simple look-ahead, the alternatives would need to be generated and the tables compared to determine a preference.

Example VI.



Partial ordering of nodes
$\{A\} > \{E\} > \{B,C,D,F\}$

Computer connectivity values

	A	B	C	D	E	F	
$i = 0$	4	1	1	1	2	1	$k = 3$*
$i = 1$	5	4	4	4	5	2	$k = 3$

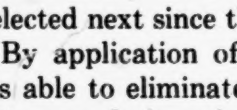
* These values used for partial ordering

In example VI the partial ordering will cause nodes A and E to be selected as the first and second nodes, respectively. At this point, however, B, C, and D are equal candidates for selection as the third, fourth, and fifth nodes, thereby giving rise to the generation of six tables. To prevent this common situation the computer detects the condition and performs a look-ahead to determine the effect of the possible choices on the next levels of the table. This look-ahead can be done since the choice cannot affect the FROM or RING CLOSURE lists. Therefore, the node values are examined and any preference implied by them is introduced. If, however, the node values are equal, the determination of a preference falls next to the line values and finally to the node and line modifications. If the choices are equal at every level then it makes no difference which is selected next since the choices give rise to identical tables. By application of this simple look-ahead the program is able to eliminate the generation of the possible alternatives and the selection from among them. In example VI, for instance, only one table will be generated instead of the six which would have been generated without the look-ahead technique.

At present the look-ahead technique is used only to deal with the case of terminal atoms. The technique could be extended, however, without affecting the ultimate choice of the unique description. Determination of whether this extension is economically required will be made on the basis of operating experience in the coming months, but at this point it seems unlikely to be necessary.

The last technique for reducing the number of tables generated is the provision to recall, under certain conditions, preferences detected during the generation process. Because of the nature of the techniques thus far described, the size of the subset is a product function based on the number of choices arising; that is, the same preference or lack of preference is rediscovered several times.

Example VII.



For instance, in example VII the preference or lack of preference between B and D will be determined twice, once when G and I are listed third and fourth, respectively, and once when G and I are listed fourth and third, respectively. It would be more efficient to remember the preference once detected and to use this information should the same choice arise again. The problem is that the preference can be recalled only when it is independent of any previous choices; therefore, the preference can be remembered only under certain conditions. These conditions are: first, the atom from which the choice arises, atom A in example VII, must be bonded to exactly three other nonhydrogen atoms, two of which are involved in the choice; and, second, the bond not involved in the choice, bond A–F in example VII, must not be part of a cycle.

If during the generation process a point of choice is reached which meets the conditions cited, the computer program divides the graph into two subgraphs by removing the bond which is not involved in the choice. In example VII the bond between A and F is temporarily removed. By virtue of the fact that the removed bond is not a member of a cycle (specified condition), the result of this removal is to divide the graph into two subgraphs. The program then operates on the subgraph involved in the choice so as to determine a preference or lack of preference between the two choices. This preference is determined by generating the set of tables which arises from the choices in the subgraph. Once such a preference is determined the graph is restored, and the preference is recalled when the same choice arises again. If there is no preference between the two choices, then it makes no difference which is selected since they are indistinguishable. In this case, the preference will be made arbitrarily and reused should the opportunity arise again.

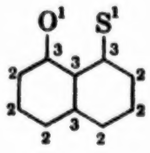
Of the several methods employed to reduce the number of tables generated the two most significant are (1) the partial ordering of the nodes by the computed "connectivity values" and (2) the rules for numbering the nodes for table generation. Together these two methods complemented by the other techniques reduce what would otherwise be a devastatingly time consuming task to one which requires only a trivial amount of time.

In order to demonstrate the presumed advantages of the techniques described, they were programmed for an IBM 1410 Data Processing System. Over 25,000 chemical structures from CAS files, selected solely on the basis of immediate availability, were processed. The description of these structures and the statistics resulting from this test are shown in Appendix II. Based on these statistics and the published timings of other techniques which have been described in the literature, it appears that the present technique offers significant economic advantage over other methods for accomplishing the same end.

APPENDIX I

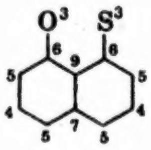
Example I.

$i = 0$



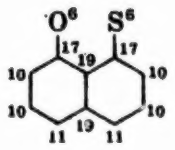
connectivity values
1, 2, 3
$k = 3$

$i = 1$



connectivity values
3, 4, 5, 6, 7, 9
$k = 6$

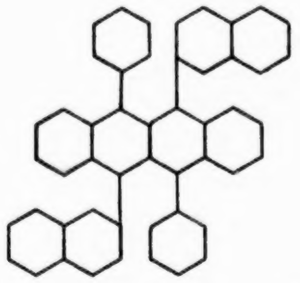
$i = 2$



connectivity values
6, 10, 11, 17, 19
$k = 5$

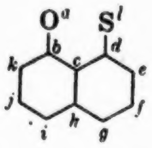
In this example the iterative process will be terminated after two iterations and the values assigned after the first will be used to introduce the partial ordering. In this example the subset of tables will consist of exactly four tables.

Example II.



	Connectivity values	Value of k	Size of the subset, tables
$i = 0$	2, 3	2	14,592
$i = 1$	4, 5, 6, 7, 8, 9	6	160
$i = 2$	8, 9, 11, 12, 14, 17, 18, 19, 20, 22, 24, 27	12	32
$i = 3$	18, 19, 20, 21, 26, 27, 28, 29, 31, 32, 38, 41, 42, 46, 50, 58, 68, 69, 75	19	8
$i = 4$	will also yield a value for k of 19; thus the process terminates and the values at $i = 3$ will be used to introduce the partial ordering of the nodes.		

Example III.



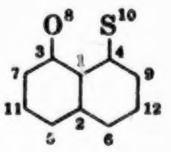
Using the connectivity values shown in example I of this appendix a partial ordering among the nodes is introduced.

$$\{c\} > \{h\} > \{b, d\} > \{k, e, i, g\} > \{j, f\}$$

Using this partial ordering the possible tables are generated and compared giving preference for lower numbering to the node or nodes which are between the leftmost pair of braces at each point of choice. For this example four tables must be generated and the unique table selected from among the set.

The preferred numbering and the corresponding unique table are shown below:

1	C	–	–
2	C	1	1
3	C	1	1
4	C	1	1
5	C	2	1
6	C	2	1
7	C	3	1
8	O	3	1
9	C	4	1
10	S	4	1
11	C	5	1
12	C	6	1
Rings	7	11	1
	9	12	1



In computer storage the unique table appears as follows:

From list	Ring closure
001001001002002003003004004005006	007011009012

Node values	Line values
CCCCCCCOCSCC	1111111111111

APPENDIX II

In order to test the presumed economic advantages of the technique described in this paper, over 25,000 chemical structures were selected from the CAS files. These structures were selected solely on the basis of immediate availability and consisted of the following:

A	The "Ring Index" structures including the first supplement	9,568
B	A CAS File of commercial compounds	7,154
C	The structures from Lange's "Handbook"	4,596
D	The CAS File of compounds containing only carbon, hydrogen and sulfur	4,287
	Total	25,605

The following is a table of statistics resulting from the testing of these techniques using the file described above:

A	Sample size	25,605 structures
B	Total 1401 computer time for the generation of the unique description	4.93 hr.
C	Average number of compounds per minute for the generation of the unique description	92.8/min.
D	Average cost per compound for the generation of the unique description	2.2 cents
E	Average number of tables generated per compound	4.3

APPENDIX E

The Computer-Based Subject Index Support System at Chemical Abstracts Service

by

D. J. Whittingham, F. R. Wetsel, and H. L. Morgan

[Reprinted from the Journal of Chemical Documentation, 6, 230 (1966).]
Copyright 1966 by the American Chemical Society and reprinted by permission of the copyright owner.

The Computer-Based Subject Index Support System at Chemical Abstracts Service[*]

D. J. WHITTINGHAM, F. R. WETSEL, and H. L. MORGAN[**]
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received October 3, 1966

This paper describes a computer-based input system which reduces or eliminates many repetitive operations. This system reduces and conserves the over-all human effort required for input of structural, nomenclature, and bibliographic data while simultaneously improving the efficiency of the registration operation and increasing the reliability of the stored data.

Since early 1965, Chemical Abstracts Service has been developing an experimental computer-based. Chemical Compound Registry System which is being supported by the National Science Foundation (NSF), the Department of Defense, the Food and Drug Administration, the National Institutes of Health, and the National Library of Medicine through a contract with NSF.

The data processed for this computer-based system involve structures, names, and references for compounds, primarily those processed in the current volumes of *Chemical Abstracts* (*CA*). Thus, each six months, the Chemical Compound Registry System receives data on about 200,000 compounds which are processed at CAS for *Chemical Abstracts*, and the *Ring Index* and its Supplements. Of these 200,000 compounds, about 163,000 have been reported previously in the literature, and have thus been processed previously at CAS. The remaining 37,000 compounds are newly reported chemical entities.

---

[*] Presented before the Division of Chemical Literature, 152nd National Meeting of the American Chemical Society, New York, N. Y., Sept. 15, 1966.

[**] Present address: IBM Corp., 1000 Westchester Ave., Harrison, N. Y.

Without the computer-based support to be described in this paper, the registration of previously reported compounds for which new data continue to be documented in the literature would involve many highly repetitive operations—drawing of structures, calculation of molecular formulas, identification of ring systems, the derivation of systematic nomenclature, keyboarding, and editing, to name a few.

This input system is an integral part of the experimental CAS Chemical Compound Registry System which has been the topic of several recent papers (*1*), and will not be discussed in detail here. However, a brief review of the component parts of the Registry System is in order to clarify its relationship to the computer-based input system. The principal components of the Chemical Compound Registry System are three major computer files of compound-oriented data. The connecting link for these three files is the permanent computer address for each compound—the Registry Number. The three files are:

1. The Structure File, which contains unique descriptions of the structural formulas including stereochemistry and isotopic labeling. The input to this file is subjected to an elaborate computer editing program, described by Leiter and Morgan (*2*).
2. The Nomenclature File, which contains all of the names for a given compound available in a variety of sources. These names are coded as to type—*e.g.*, preferred *CA* index name and trade name. Presently, editing for the Nomenclature File is done mainly by chemists during the production of the *CA* indexes. The existence of the Nomenclature File makes possible the printing of a variety of special indexes—*e.g.*, an index of trade names versus the *CA* preferred index names.
3. The Bibliography File, which contains references to the *CA* abstracts. By use of other existing computer files these references can be coordinated with the corresponding original journal references.

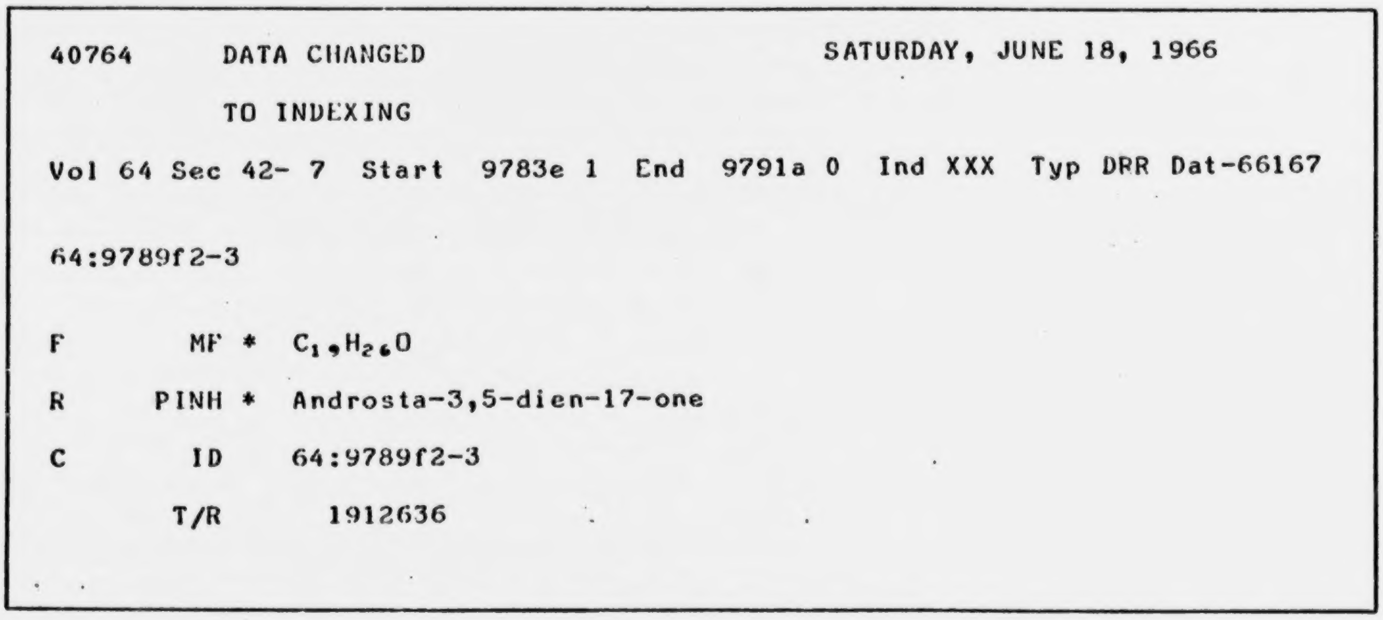
These files of the experimental CAS Compound Registry System are reviewed for accuracy through the efforts of CAS chemists and clerical personnel, aided by the computer-based input system described in the following pages. Two types of computer support are provided through this system, one based on the Structure File, and one based on the Nomenclature File.

SUPPORT THROUGH THE STRUCTURE FILE

The first set of procedures, based on the Registry System Structure File, is applied to compounds registered from those sections of *CA* reporting the highest percentage of new compounds—namely, the synthetic organic sections.

Compounds selected for registration from these sections have their structures drawn and molecular formulas calculated by CAS professional staff. The structural formulas are then clerically processed for registration. In this operation, all nonhydrogen atoms in the structural formulas are numbered in any convenient sequence. Then the atoms and bonds are keyboarded in tabular form for input to the computer. In the same operation the *CA* reference, any trivial names, and the calculated molecular formula are keyboarded for computer input (*3*). The structure is then registered, a process in which a computer program compares the structural information input with the data on file for all other structures. Registration results in one of two situations.

The first, a "hit," means that the compound has been registered previously and thus is already on file. In such cases, the computer program retrieves from the Registry Files the molecular formula, the edited and correctly formatted index name(s), and the Registry Number for the compound. This information, together with the originally keyboarded *CA* reference for the current volume, is printed by the computer on a data sheet (Figure 1) and sent to a *CA* chemist for review. All of the information on this sheet is reviewed for accuracy and corrected for discrepancy as necessary.



```
40764     DATA CHANGED                              SATURDAY, JUNE 18, 1966

          TO INDEXING

Vol 64 Sec 42- 7  Start  9783e 1  End  9791a 0  Ind XXX  Typ DRR Dat-66167

64:9789f2-3

F         MF *  C19H26O

R       PINH *  Androsta-3,5-dien-17-one

C         ID    64:9789f2-3

         T/R      1912636
```

Figure 1. Structure file "hit" computer proof sheet.

Through the information retrieved from the Registry File, the input system has reduced the effort required for registration. That is, compounds for which hits result do not require naming, and the names do not require keyboarding, editing, or re-entry into the computer.

Additionally, if the compound contains a ring system, the computer programs will determine whether the ring system is new or has been registered previously. (Records for some 18,000 ring systems are on file as a result of the registration of the *Ring Index* and its Supplements.) Where there is a hit, the computer prints a data sheet (Figure 2) giving the Registry Number, Ring Index Number, and molecular formula of the ring system. This saves the effort of renaming and reregistering the ring system.

The second situation that may obtain as a result of registration is "no hit," meaning that the compound is new to the Registry System. In such cases, the computer prints a data sheet (Figure 3) containing the molecular formula and the *CA* reference for the current volume (that is, the data that were input to the computer for registration). These data are proofread to assure that they have entered the computer accurately, and then CAS professional staff derive the *CA* index name(s), which are keyboarded and entered into the Registry Files.

SUPPORT THROUGH THE NOMENCLATURE FILE

The set of input procedures based on the Registry System Name File applies to sections of *CA* other than the synthetic organic sections. These sections contain a high percentage of compounds reported previously in the literature.

CAS staff concerned with the selection of compounds to be registered dictate the available systematic or trivial names of the selected compounds and the corresponding *CA* references. This information is then keyboarded and input to the computer, where each name is compared with the names already filed in the Nomenclature File. As with Structure File support, one of two situations obtains when a systematic or trivial name is compared with names on file in the computer.

The first, a "hit," means that the input name is already on file and therefore that the compound has been registered previously. In these cases, the computer program retrieves from the files the molecular formula, the edited and correctly formatted *CA* index name(s), and the Registry Number for the compound. This information is printed by the computer together with the name and *CA* reference dictated at the time of the compound selection. Figure 4 illustrates such a computer-printed proof sheet.

Through this procedure, the computer input system has again reduced the effort required for registration. For compounds for which hits result, the structural formulas do not have to be drawn, the molecular formulas do not have to be calculated, and the preferred *CA* index names do not have to be generated, keyboarded, edited, or re-entered into the computer.

The second situation that may obtain after name input is "no hit," meaning that the name is not on file. For such situations, structural formulas must be drawn and the molecular formulas calculated. The registration process then proceeds, with further support obtained as previously described under Support Through the Structure File.

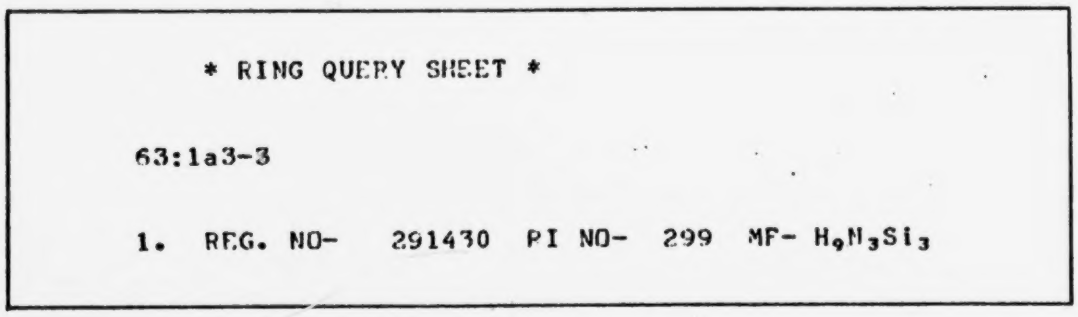
Even in the "no hit" situation, however, the name and associated data have been added to the computer record. The next time the same name is encountered in *CA* indexing, computer support through the Name File will be possible.

SOME SAVINGS OF CLERICAL EFFORT

Two areas where the computer-based support system has effected considerable savings in the use of clerical effort involve the single keyboarding of preferred index names and *CA* references, and the use of keyboarding "shortcuts" that save keystrokes during data input.

CA index operations currently use as manuscript 3 × 5 cards containing one entry per card. For compound entries, the name of the compound and the *CA* reference appear on two cards, one for the Subject Index, and one for the Formula Index. Before the computer-based input system was developed, these 3 × 5 cards were typed directly. In addition the registration operation then required another keyboarding of many compound names and *CA* references. These operations are now all combined into a single keyboarding of data for entry into the computer, which then delivers the index cards and records the appropriate information in the Registry. The total keyboarding effort has been reduced by an estimated 25% as a result of this procedure.

The use of keyboarding shortcuts has also been made possible through the use of the computer input system. Under the old system, the typists were never afforded the possibility of shortcuts. For example, the registration

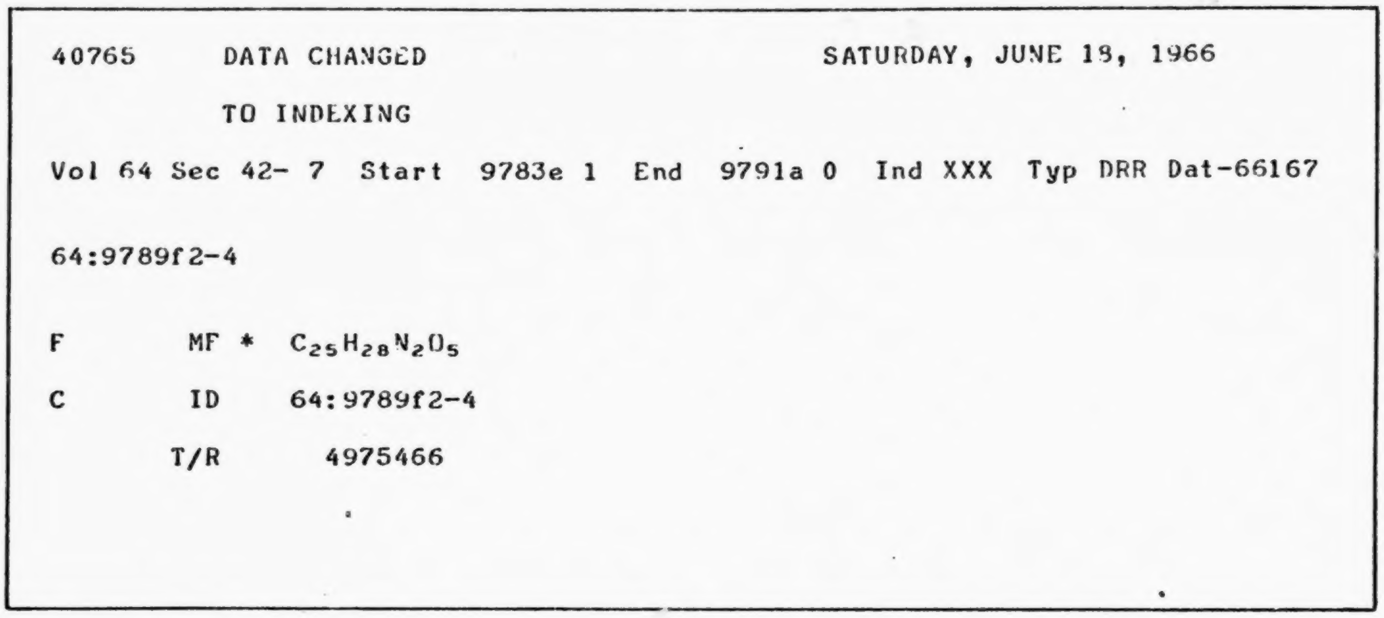


```
     * RING QUERY SHEET *

63:1a3-3

1.  REG. NO-   291430  RI NO-  299  MF- H9N3Si3
```

Figure 2. Computer-produced ring query sheet.



```
40765      DATA CHANGED                             SATURDAY, JUNE 18, 1966

           TO INDEXING

Vol 64 Sec 42- 7  Start  9783e 1  End  9791a 0  Ind XXX  Typ DRR Dat-66167


64:9789f2-4


F          MF *  C25H28N2O5

C          ID    64:9789f2-4

        T/R        4975466
```

Figure 3. Structure file "no hit" computer proof sheet.

of six different esters of 2,4,6-triiodobenzoic acid required the keyboarding of the parent acid name six times.

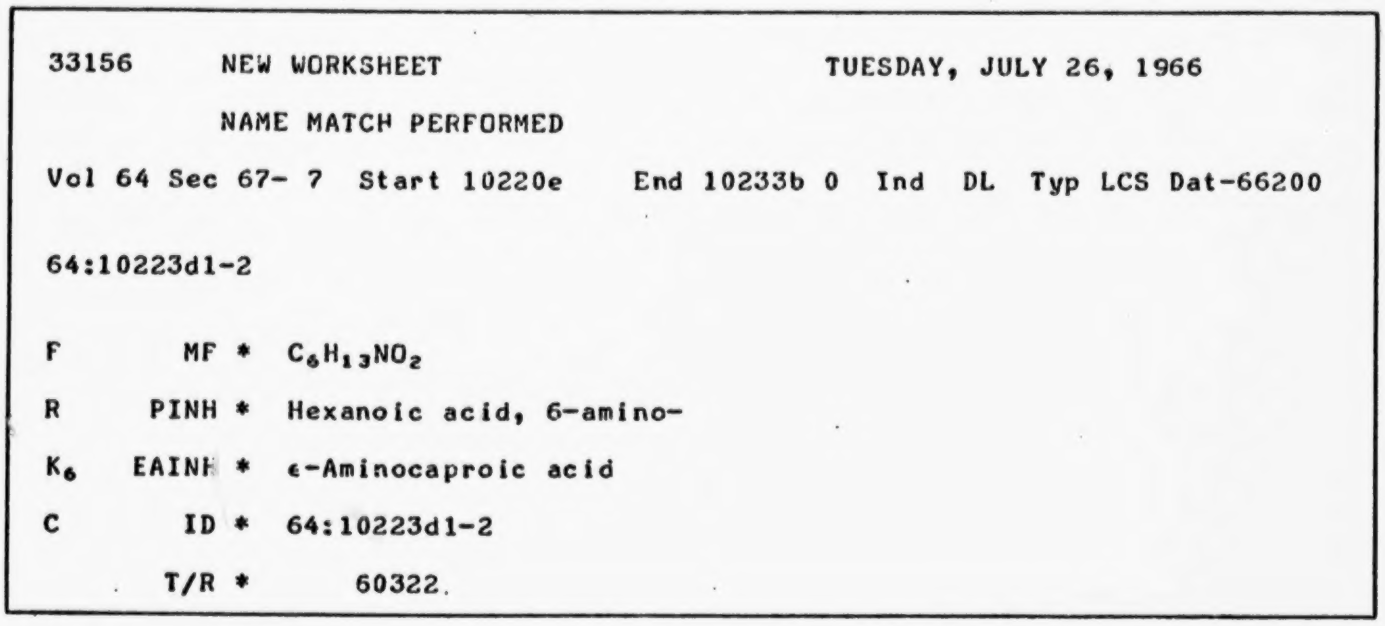
In the computer-based input system, however, the typists now keyboard the complete name of the parent acid once. For the five remaining esters, a two-character "ditto code" is typed instead of the parent acid name (Table I). The ditto code instructs the computer to print out the parent acid name keyboarded for the first entry. Note that the ditto code is fully expanded by the computer and thus is not carried in the permanent files. Therefore, it makes no difference in the stored data whether or not the typist uses the shortcut.

Two other ditto features are also used. One is used to repeat a name's modification (that portion of the name that appears in light-face type in the *CA* Subject Index) from one entry to another. The second, used in registering alphabetized lists of names, repeats that portion of a name up to and including the first comma followed by a space—*e.g.*, the comma of inversion. We estimate that 42,000 keystrokes are saved by these three features every work day; this is equivalent to 5% of the total keyboarding effort.

PROJECTED SUPPORT IN PRINTED ISSUE, VOLUME, AND COLLECTIVE INDEXES

In 1967, CAS expects to install a much more advanced computer support system, which will be directed primarily at support of index operations rather than Registry operations. Through this system, CAS will eliminate the use of the 3 × 5 manuscript cards, and, instead, produce "camera-ready" copy from the computer for final proof and printing operations. This will result in greater computer support for the indexing operations. For example:

1. Alphabetizing of the entries for an index will be performed as a computer operation.
2. The merging of edited index volumes to collective indexes will also be performed as a computer operation.



```
33156      NEW WORKSHEET                            TUESDAY, JULY 26, 1966

           NAME MATCH PERFORMED

Vol 64 Sec 67- 7  Start 10220e    End 10233b 0  Ind  DL  Typ LCS Dat-66200


64:10223d1-2


F          MF *  C6H13NO2

R        PINH *  Hexanoic acid, 6-amino-

K6      EAINH *  ε-Aminocaproic acid

C          ID *  64:10223d1-2

          T/R *      60322.
```

Figure 4. Name file "hit" computer proof sheet.

Table I. Example Use of "Ditto Code" to Save Keystrokes

Typed as	Computer-printed as
Benzoic acid, 2,4,6-triiodo-, ethyl ester	Benzoic acid, 2,4,6-triiodo-, ethyl ester
Ipmethyl ester	Benzoic acid, 2,4,6-triiodo-, methyl ester
Ippropyl ester	Benzoic acid, 2,4,6-triiodo-, propyl ester
Ipisopropyl ester	Benzoic acid, 2,4,6-triiodo-, isopropyl ester
Ipbutyl ester	Benzoic acid, 2,4,6-triiodo-, butyl ester
Ip3-nitropropyl ester	Benzoic acid, 2,4,6-triiodo-, 3-nitropropyl ester

Further, new indexes and a data base for searching property, reaction, and use information from the computer-based Subject Index Support System will become a reality.

Our users will receive the first products of this completely mechanized system in 1967 in the form of the computer-composed volume Author and Formula Indexes. Note, however, that the Registry Support System described in this paper is in operation at CAS at this time.

APPENDIX

Explanation of Terms Used on Computer-Produced Data Sheets

The data sheets produced by the computer as described in the foregoing text carry several types of information for use in entering compounds into the CAS Chemical Compound Registry System. The sheets are printed in worksets grouped by *CA* sections and by column fractions, and within a given workset, each compound is represented by a single data sheet. The following is an explanation of the terms used on the data sheets in Figures 1, 3, and 4 of this paper.

The heading of each sheet includes a sequential number assigned by the computer, an identification of the type of sheet—*e.g.*, "new worksheet"—and the day and date on which the data were processed by computer.

The following data are keyboarded once for each workset and are then automatically printed by the computer on each applicable data sheet.

Vol	The volume of *CA* from which the data were obtained.
Sec	The section and issue number of *CA* from which the data were obtained.
Start-End	The limits (expressed as column numbers, letter fractions, and abstract numbers of the *CA* issue) between which the data were obtained.
Ind	The chemist, identified by initials, who dictated the data. (The examples use XXX as the indexer's initials.)
Typ	The typist, identified by initials, who keyboarded the data.
Dat	The date the data were keyboarded.

The following are codes for major fields keyboarded for each data set:

F	Formula.
R	Preferred *CA* index name.
N	Added *CA* index name.
K_{1-5}	Extra added *CA* index name.
K_{6-9}	Fields in which systematic, trade, or trivial names are input for Index Support through the Nomenclature File.
C	Identification or reference.

The following are codes for subfields printed automatically by the computer:

MF	Molecular formula.
PINH	Preferred index name heading, that portion of the index entry that appears in bold-face type in the subject indexes.
EAINH	Extra added index name heading.
ID	The *CA* reference including column number, letter fraction, abstract number, and compound number. The latter two numbers are for internal computer use only.
T/R	The temporary identification number (T) or the Registry Number (R) of the compound. Temporary identification numbers are used in initial processing until a Registry Number is assigned. The Registry Number becomes a compound's permanent identification in the CAS Chemical Compound Registry System.

LITERATURE CITED

(1) See, for example, Leiter, D. P., Jr., Morgan, H. L., Stobaugh, R. E., *J. Chem. Doc.* **5**, 238 (1965).

(2) Leiter, D. P., Jr., Morgan, H. L., "Quality Control and Auditing Procedures in the Chemical Abstracts Service Compound Registry," *ibid.*, **6**, 226 (1966).

(3) These clerical operations are described more fully in Leiter *et al.*, *ibid.*, **5**, 240-1 (1965).

APPENDIX F

Improvements in Structure Registry Effected with the Redesign and Reprogramming for IBM 360 Computers

Extracted from "System and Program Documentation
for the Chemical Abstracts Service Registry System"
Copyright 1968 by the American Chemical Society

I. Systems Improvements

1. Standard File Formats

Standard formats have been established for each item of data in each file. This provides several advantages:

a) Each data element is well-defined and has consistent representation throughout the system.

b) Standard formats provide interface continuity between files - - a given data element has the same format in all files.

c) Because data elements have standard formats, standard input and output programs need be written only once, then reused as needed. These standard input/output routines are included in the CAS standard subroutine library.

d) Eleven of the 12 data fields on the Registry Structure Master File are variable in length. Therefore, the system will adapt to future changes in data length without requiring program changes.

e) Standard formats simplify searching and can therefore improve search retrieval.

2. Improved 360 Hardware/Software Capabilities

The reprogrammed systems will take advantage of improved equipment and software capabilities offered by the 360 computers.

a) Processing is speedier because of faster core access times.

b) The ability to process data one-half byte at a time permits smaller units of data to be manipulated and compacts the file further.

3. Modular Programming

The redesigned system will be written as a set of program modules, each designed to perform a specific system function. This offers several advantages:

a) It increases the adaptability and flexibility of the system because modules can be changed without affecting the rest of the system.

b) It will be easier to modify programs, since systems personnel will work with smaller program "pieces".

c) Modular programs make it easier to utilize subroutines from the library.

d) Modular programming allows for future expansion -- for example, a module for interfacing the Registry with the Substructure Search System.

II. Technical Improvements

1. Tautomers

The new system provides computer programmed identification of unique compounds that can be represented by two different, but equally valid, structural diagrams.

A generalized representation is:

$$M{=}Q{-}(Q{=}Q)_n{-}ZH \rightleftarrows H{-}M{-}(Q{=}Q)_n{-}Q{=}Z$$

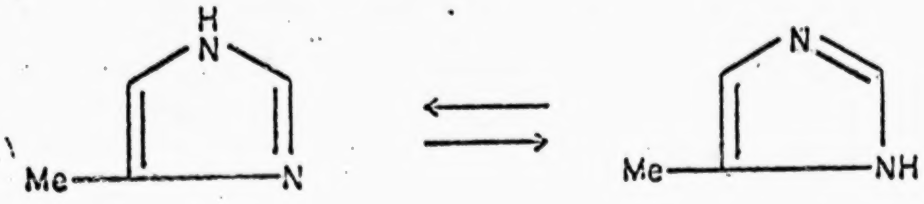
where M, Q, and Z are combinations of C, N, P, As, O, S, Se, and Te (including abnormal mass analogs); $n \geqq 0$ (integral); = is a double bond; - is a single bond; and H must be present as shown.

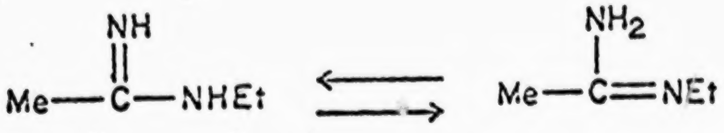
The following are examples of the types of structures involved:

a) General representation

$M{=}Q{=}(Q{=}Q)_n{-}Z{-}H$
M=Z=N(trivalent)
Q=C

n is limited to zero in the initial system; the value can be extended in subsequent versions (1,2,...) if justified. Analogous structures with P or As may be found to exhibit tautomerism; provision is made for such an extension if necessary.



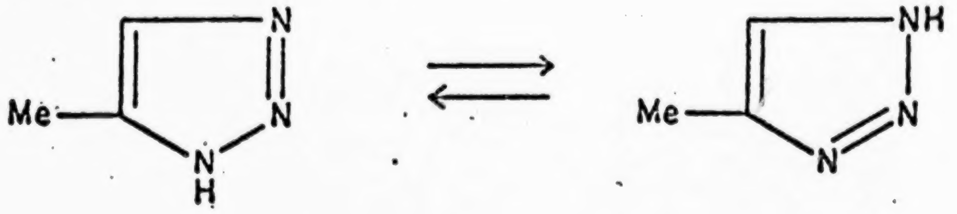


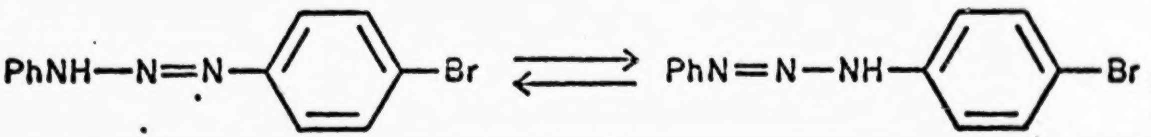
b) General expression

$M{=}Q{=}(Q{=}Q)_n{-}Z{-}H$

M=Q=Z=N

As in a, n is initially limited to zero and provision is made for future substitution of P or As for N.





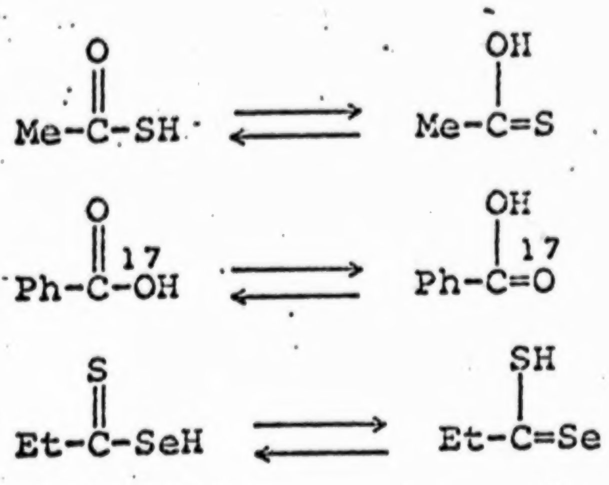
c) General expression

$M=Q-(Q=Q)_n-Z-H$

M=Z=O,S,Se,Te (and abnormal mass analogs)

Q=C

n=0



$$\mathrm{Me{-}\overset{\overset{O}{\|}}{C}{-}SH \rightleftharpoons Me{-}\overset{\overset{OH}{|}}{C}{=}S}$$

$$\mathrm{Ph{-}\overset{\overset{O}{\|}}{C}{-}{}^{17}OH \rightleftharpoons Ph{-}\overset{\overset{OH}{|}}{C}{=}{}^{17}O}$$

$$\mathrm{Et{-}\overset{\overset{S}{\|}}{C}{-}SeH \rightleftharpoons Et{-}\overset{\overset{SH}{|}}{C}{=}Se}$$

d) General expression

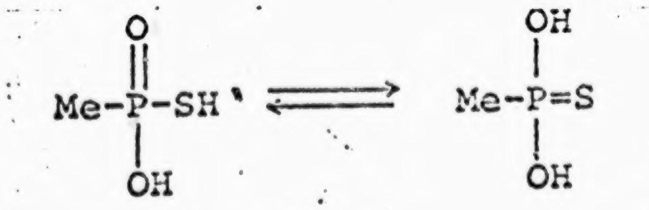
$M{=}Q{-}(Q{=}Q)_n{-}Z{-}H$

$M{=}Z{=}O, S, Se, Te$ (and abnormal mass analogs)

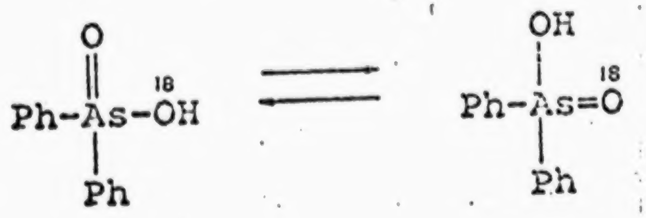
$Q{=}N, P, As, Sb$ (tri- or pentavalent),

S, Se, Te (tetra- or hexavalent), Cl, Br, I

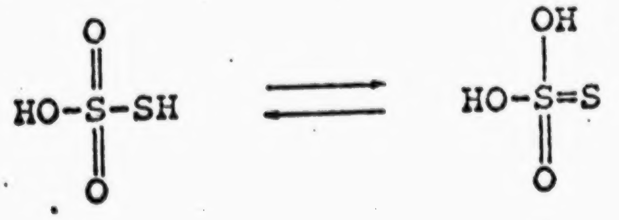
$n{=}0$



$$\mathrm{Me{-}\underset{\underset{OH}{|}}{\overset{\overset{O}{\|}}{P}}{-}SH} \rightleftharpoons \mathrm{Me{-}\underset{\underset{OH}{|}}{\overset{\overset{OH}{|}}{P}}{=}S}$$



$$\mathrm{Ph{-}\underset{\underset{Ph}{|}}{\overset{\overset{O}{\|}}{As}}{-}{}^{18}OH} \rightleftharpoons \mathrm{Ph{-}\underset{\underset{Ph}{|}}{\overset{\overset{OH}{|}}{As}}{=}{}^{18}O}$$



$$\mathrm{HO{-}\underset{\underset{O}{\|}}{\overset{\overset{O}{\|}}{S}}{-}SH} \rightleftharpoons \mathrm{HO{-}\underset{\underset{O}{\|}}{\overset{\overset{OH}{|}}{S}}{=}S}$$

2. Improved Handling of Ring Bonds

The new system incorporates improved procedures for handling cyclic systems. The paths to identify rings are traced starting with the ring closure pairs, thus eliminating the duplicate tracing that resulted from a trial-and-error basis. This technique provides improved efficiency as to type - - ring, chain or alternating single-double.

3. Automatic Editing of Text Descriptors

The 360 system includes editing routines for text (stereochemical) descriptors which simplify structure problem resolution by eliminating part of the human involvement in the process. The amount of chemist time for such resolution will be reduced by about 35%.

A table of standard, valid text descriptors including synonyms is used in a computer edit of this feature of the connection table. The program checks for the presence (exact match) of the input text descriptor in this table; rejects the table if the input text descriptor is not valid (unless an override is coded); corrects certain descriptors (for the purpose of chemical checking because of ambiguity, marked *** in the table); and performs automatic resolution in ties involving valid unambiguous descriptors.

Text descriptors have been developed for the main body of steroids, terpenes, alkaloids, and carbohydrates based upon an alphabetic term or base name, representing a basic parent structure with implied stereochemistry at specified positions. This base name is the name of the parent structure or a term closely related to it.

For steroids, terpenes, and alkaloids, substituents attached to the parent at a potentially asymmetric atom require a locant or locants preceding the base name. This consists of a numerical

2. Improved Handling of Ring Bonds

The new system will apply improved procedures for handling cyclic systems. The paths to identify rings are traced starting with the ring closure pairs, thus eliminating the duplicate tracing that resulted from a trial-and-error basis. This technique provides improved efficiency as to type -- ring, chain or alternating single-double.

3. Automatic Editing of Text Descriptors

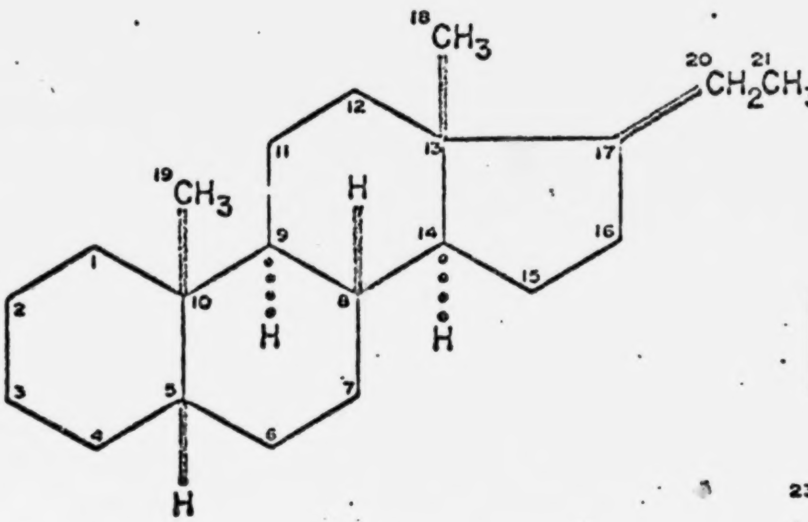
The 360 system will provide editing routines for text (stereochemical) descriptors which will simplify structure problem resolution by eliminating part of the human involvement in the process. The amount of chemist time for such resolution will be reduced by about 35%.

A table of standard, valid text descriptors including synonyms is to be used in a computer edit of this feature of the connection table. The program is to check for the presence (exact match) of the input text descriptor in this table; to reject the table if the input text descriptor is not valid (unless an overide is coded); to correct certain descriptors (for the purpose of chemical checking because of ambiguity, marked *** in the table); and to perform automatic resolution in ties involving valid unambiguous descriptors.

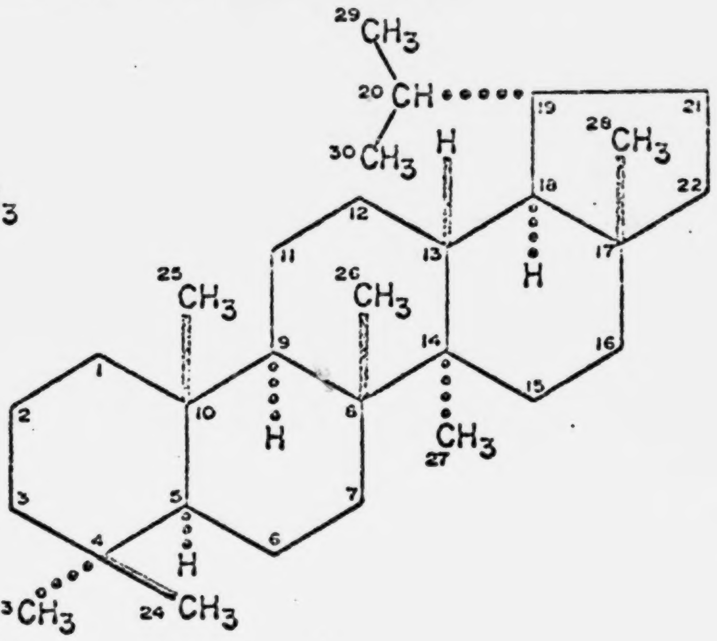
Text descriptors have been developed for the main body of steroids, terpenes, alkaloids, and carbohydrates based upon an alphabetic term or base name, representing a basic parent structure with implied stereochemistry at specified positions. This base name is the name of the parent structure or a term closely related to it.

For steroids, terpenes, and alkaloids, substituents attached to the parent at a potentially asymmetric atom require a locant or locants preceding the base name. This consists of a numerical locant for the substituent, based on the CA accepted numbering for the parent structure, followed by one of three possible letters showing the configuration of the attached substituent. The α (alpha) configuration, representing a projection below the plane of the paper, is shown in the illustrative examples by a broken line and represented by A as a locant. The β (beta) configuration, representing a projection by a heavy solid line and represented by B as a locant. The ξ (Xi) or unknown configuration is shown in the examples by a wavy line and represented by X as a locant.

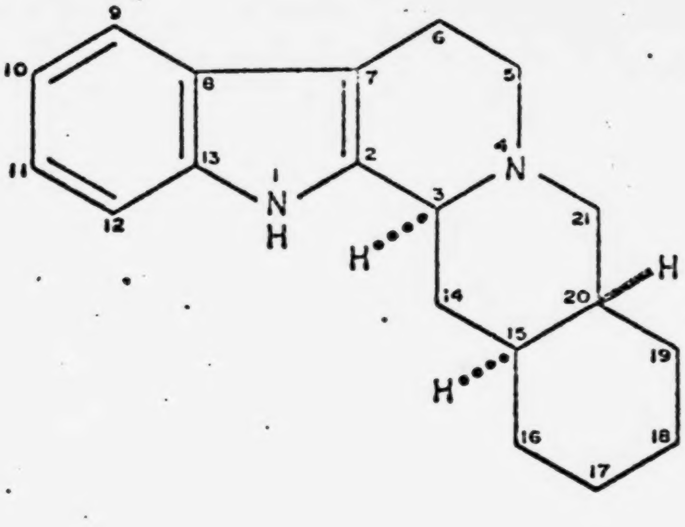
Examples are given, of a steroid (PREGN), terpene (LUPANE), and alkaloid (YOHIMBAN), base structure and corresponding base name. Standard numbering and stereochemistry are also shown.



PREGN

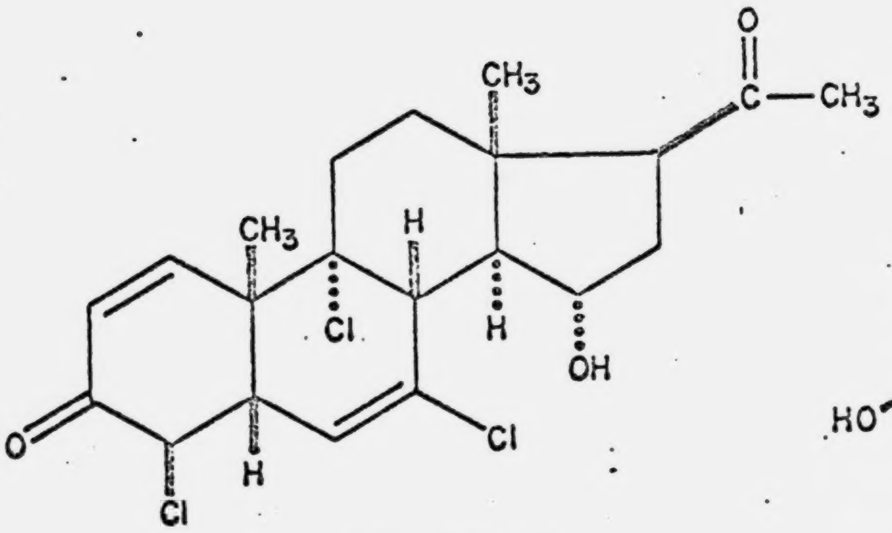


LUPANE

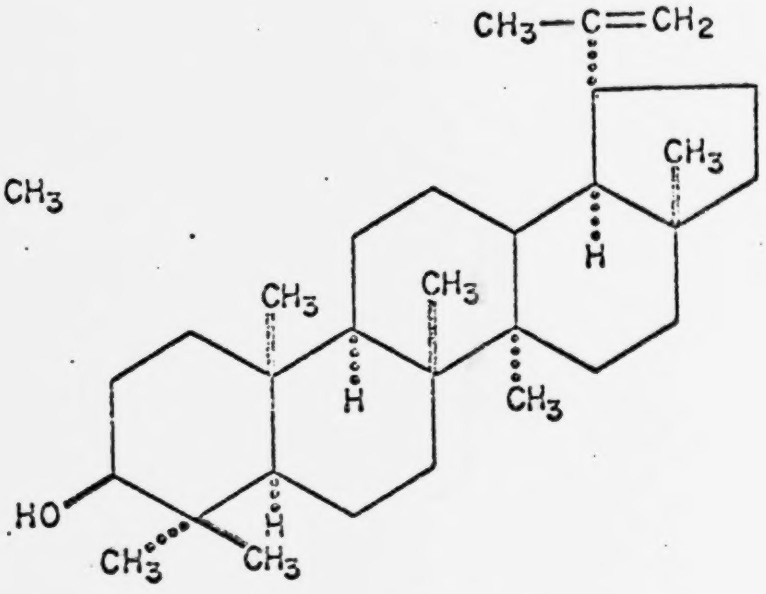


YOHIMBAN

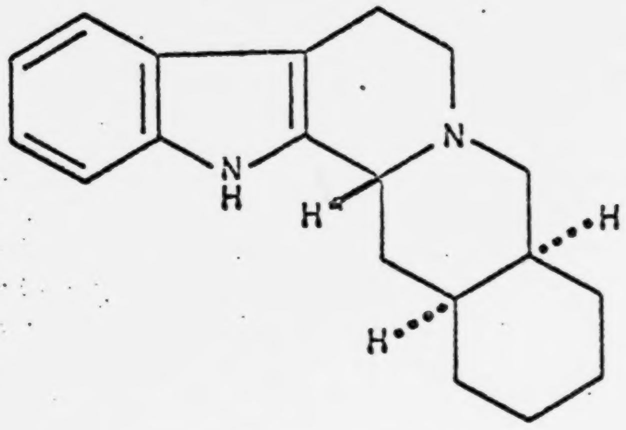
Examples of the use of these text descriptors follow:



4B,15A-PREGN

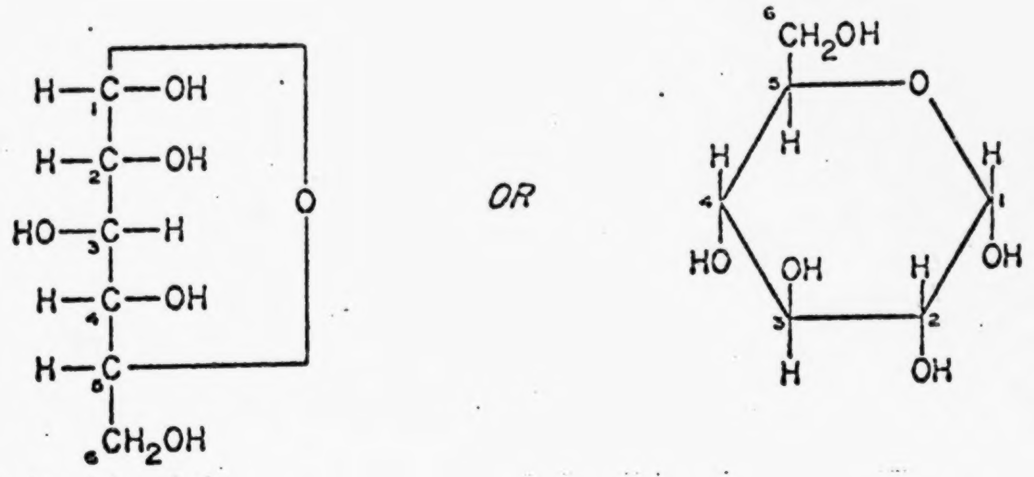


3B-LUPANE



3B,20A-YOHIMBAN

For carbohydrates, text descriptors are based primarily on the name of the compound. While the text descriptors can be generated from the structure only, an adequate name is usually found in connection with the structure. The systematic names of monosaccharides are derived from the trivial names of the sugars themselves, e.g., glucose, mannose, ribose. The word roots of these names indicate the stereochemistry of some of the hydroxyl groups (or derivatives), and the anomeric prefix and configurational prefix complete the definition of stereochemistry in the compound. For illustration, in the name α-D-glucopyranose, α is the anomeric prefix, D the configurational prefix, and gluco the word root. The structural diagram is commonly as follows:



α defines the stereoisomerism at C-1, D at C-5, and gluco at C-2, and C-4. Thus the combination α-D-gluco, or as written for the text descriptor, A-D-GLUCO defines the stereochemistry of the compound.

4. Improved Handling of Coordination Compounds

Detailed structuring conventions for coordination compounds have been established to assure unique identification of each compound and increase the level of detail available for substructure searching. The structure conventions are designed to provide the two important values associated with the central atom of the coordination compound:

Coordination Number: the sum of the bond lines of attachment to the atom is equal to the coordination number.

Oxidation State (Stock Number, valence): the number of charges on the central atom (positive, negative, or zero) is equal to the oxidation state.

In connection with these two points, appropriate computer edit of bonds and number (and kind) of charges for given elements is provided. For each element treated there are allowable charges and for each of these there is one or more permitted number of bonds. The table (Appendix B) of element symbols with permitted charges and corresponding bond lines is used in this edit.

> N. B. Whereas "coordination number" and "oxidation state" are used for input edit and are a part of the machine record, these numbers can be eliminated as a part of the direct structural output being developed for the System.

5. Additional Edit Checks for Abnormal Mass Citations

Unlike the 7010 system, which did not check abnormal mass citations, the 360 system will check such citations against a stored list of acceptable values for initial mass checking, these being the most common cited in chemical and biochemical texts, reference works, and abstracts. The elements and mass values are given in the table. Unless the input structure contains a permissible mass, the structure will be rejected.

6. Editing Checks for Stock Numbers

For a selected list of multivalent metals, a list of acceptable oxidation states (represented by Stock numbers) has been established. The structures concerned are the metal salts of acidic compounds, represented in the structure file as "disconnected" structures. A multivalent metal must have an associated Stock number in this situation. Monovalent metals do not require a stock number. The 360 system will check input Stock numbers against a table before accepting a structure.

7. Registration of Larger, More Complex Molecules

Largely because of the increased speed and capacity of the 360 computers, several arbitrary limitations of the size and complexity of molecules can be removed.

a) The maximum number of nonhydrogen atoms per compound has been increased from 150 to 254 for machine registration.

b) The maximum permissible number of nonhydrogen attachments to any one atom has been increased from six to 15.

c) The maximum permissible number of paths traced during unique table generation has been increased from a kind of 10^3 to the number traced in a time limit of 2 minutes and 10 seconds. This will permit highly symmetrical molecules to be registered. This is particularly useful for coordination compounds, in which six, seven, or eight attachments to one metal atom may give rise to a more complex type of symmetry than exists for the usual, organic compounds with up to four attachments to carbon atoms.

III. User Oriented Improvements

While all improvements to the 360 Registry will improve the value, reliability and speed of the system for its users, several changes are specifically user-oriented.

1. Structure Match Without Registration

Structure match against the Registry File will be possible without registration. That is, compounds can be matched without adding them to the file. This will be a significant advantage to users with confidential or proprietary compounds.

2. Molecular Formula Processing

Unlike the 7010 system, which checked the molecular formula early in processing and then dropped the formula until nearly the end of processing, the 360 system carries the formula through all structure processing steps. This permits the user to make use of a portion of the CAS registration processing, that is, exit from the system at one of several possible points prior to registration, and maintain a complete record. This record contains the molecular formula in the 360 system, whereas it did not in the 7010 system.

3. Substructure Search Improvements

The new system will explicitly record the number of hydrogen atoms bonded to each noncarbon atom in a compound. (Previously, only the total number of hydrogen atoms in the molecule had been recorded). This change will provide greater structural detail for substructure searching.

Valid Text Descriptor Terms

ALL
ALLO
ALLOCINNAMIC
ALTRO
AMBRANE
AMBROSANE
ANDROST
ANTI
APORPHINE
ARABINO
ATISANE
BERBINE
BRUCINE
BUF
CADINANE
CARD
CATHARANTHINE
CEPHALOSPORANIC
CEPHALOSPORIN
CEVANE
CHLORAMPHENICOL
CHLOROGENIC
CHOL
CHOLEST
CIS***

COA
CONANINE
CRINAN
CYCLOHEXIMIDE
D
DAMMARANE
DECAMER
DIELDRIN
DIISOTACTIC
DIMER
DISYNDIOTACTIC
DL
E
EMETINE
ENDO***
ENDRIN
EPHEDRINE
EPHENAMINE
EREMOPHILANE
ERGOLINE
ERGOST
ERYTHRO
ESTR
ETHAMBUTOL
EUCANINE

EUDESMANE
EXO
FACET
FOLATE
FUROST
GALACTO
GALANTHAMINE
GAMBOGIC
GAMMACERANE
GERMACRANE
GIBBANE
GLIOTOXIN
GLUCO
GLYCERO
GON
GUAIANE
GULO
HEPTAMER
HEXAMER
IDO
ISOCHLOROGENIC
ISOMORPHINAN
ISOPULEGONE
ISOTACTIC
KAURANE

*** Ambiguous descriptor requiring technical review.

L
LABDANE
LANOST
LINDANE
LUPANE
LYCORAN
LYSERGIC
LYXO
MANNO
MENTHOL
MESO
METARAMINOL
MORPHINAN
MUCO***
MYO***
NEOCHLOROGENIC
NEURAMINIC
NONAMER
NS
OCTAMER
OLEANANE
ONOCERANE
OXYTOCIN
PENICILLIN
PENTAMER
PHYLLOCLADANE
PODOCARPANE
PREGN
PSEUDOEPHEDRINE
PSEUDOTROPINE
PULEGONE
QUINIC
QUINIDINE
QUININE
R
RIBO
ROSANE
S
SCOPOLAMINE
SCYLLO***
SECURININE
SHIKIMIC
SIALIC
SOLANIDANE
SOLASODANE
SPARTEINE
SPIROST
STIG
STRYCHNINE
SYN
SYNDIOTACTIC
TALO
TETRACYCLINE
TETRAMER
THREO
TOMATIDANE
TRANS***
TRIMER
TROPANE
TROPINE
URSANE
VINCAMINE
X
XYLO
YOHIMBAN
Z
(+)***
(-)***
(±)***
*(asterisk)

*** Ambiguous descriptor requiring technical review.

Table of Charges and Bonds for Coordination Compounds

Element Symbol	Charges	Bond Lines
Ac	3+	6
Ag	1+ 2+	2,4 4
Al	3+	4,6
Am	3+	6
As	3+ 5+	4 6
Au	1+ 3+	2 4
B	3+	4
Ba	2+	4,6
Be	2+	4
Bi	3+ 5+	4 6
Bk	3+	6
Ca	2+	4,6
Cd	2+	4,6
Ce	3+ 4+	6 6
Cf	3+	6
Cm	3+	6
Co	1- 0 1+ 2+ 3+	4 4,5,6 4,5,6 4,5,6 6

Element Symbol	Charges	Bond Lines
Cr	0	6
	2+	4,6
	3+	6
	6+	8
Cs	1+	4,6
Cu	1+	2,3,4
Dy	3+	6
Er	3+	6
Es	3+	6
Eu	2+	4
	3+	6
Fe	0	5,6
	2+	4,6
	3+	4,6
Fm	3+	6
Fr	1+	4,6
Ga	1+	4,6
	3+	4,6
Gd	3+	6
Ge	2+	4
	4+	6
Hf	4+	6,6
Hg	1+	2,4
	2+	4,6
Ho	3+	6
In	1+	4
	3+	4,6
Ir	0	5,6
	1+	4
	2+	4,5
	3+	6
	4+	6

NAME
System 360 Registry
Program
Module
Macro

ID
System A015
Program
Module
Macro

TOPIC Charges and Bonds for Coordination Compounds

PAGE

Element Symbol	Charges	Bond Lines
K	1+	4,6
La	3+	6
Li	1+	4
Md	3+	6
Mg	2+	4
Mn	0	6
	1+	6
	2+	4,5,6
	3+	6
	4+	6
	6+	8
	7+	8
Mo	0	6
	2+	4
	3+	6,8
	4+	5,6,8
	5+	6,8
	6+	8
Na	1+	4,6
Nb	1-	6
	2+	6
	3+	6
	4+	6
	5+	6,7,8
Nd	3+	6
Os	0	5,6
	2+	6
	3+	6
	4+	6
	6+	8
	8+	9
P	3+	4
	5+	6
Pb	2+	4
	4+	6

CHEMICAL ABSTRACTS SERVICE
Systems Development

DOCUMENTATION MANUAL

NAME
System 360 Registry
Program
Module
Macro

ID
System A015
Program
Module
Macro

TOPIC Charges and Bonds for Coordination Compounds

PAGE

Element Symbol	Charges	Bond Lines
Pd	2+	4
	4+	6
Pm	3+	6
Pr	3+	6
	4+	6
Pt	2+	4
	4+	6
Pu	3+	6
	4+	6,8
	5+	6,8
	6+	8,10
Ra	2+	4,6
Rb	1+	4,6
Re	0	6
	3+	4,6
	4+	6
	5+	6,8
	6+	8
	7+	8,9
Rh	0	5,6
	1+	4,5
	2+	5
	3+	6
	4+	6
Ru	0	5,6
	2+	6
	3+	6
	4+	6
	6+	8
	7+	8
Sb	3+	4
	5+	6,8

CHEMICAL ABSTRACTS SERVICE
Systems Development
DOCUMENTATION MANUAL

NAME
System 360 Registry
Program
Module
Macro

ID
System A015
Program
Module
Macro

TOPIC Charges and Bonds for Coordination Compounds
PAGE

Element Symbol	Charges	Bond Lines
Sc	3+	6
Se	4+	6
Si	4+	6
Sm	3+	6
Sn	2+	4
	4+	6
Sr	2+	4,6
Ta	1-	6
	2+	6
	3+	6
	4+	6
	5+	6,7,8
Tb	3+	6
Tc	4+	6
	7+	8
Te	4+	6
Th	3+	6
	4+	8
Ti	2+	6
	3+	6
	4+	6,8
Tl	1+	4
	3+	4,6
Tm	3+	6
U	3+	6
	4+	6,8
	5+	6,8
	6+	8,10

CHEMICAL ABSTRACTS SERVICE
Systems Development

DOCUMENTATION MANUAL

NAME
System 360 Registry
Program
Module
Macro

ID
System A015
Program
Module
Macro

TOPIC Charges and Bonds for Coordination Compounds

PAGE

Element Symbol	Charges	Bond Lines
V	0	6
	2+	6
	3+	6
	4+	5,6
	5+	6
	2-	5
W	0	6
	2+	4
	3+	6,8
	4+	5,6,8
	5+	6,8
	6+	8
Y	3+	6
Yb	2+	4
	3+	6
Zr	4+	6,8

Table of Acceptable Abnormal Mass Values

Element Symbol	Acceptable Mass Values
Au	195,198,199
Br	77,79,81,82
C	11,13,14
Ca	45,47
Cl	36,38
Co	56,57,58,60
Cr	51
I	124,125,129,131,132
K	42
N	15
Na	24
O	17,18
P	32
S	35
Sr	90

Table of Acceptable Stock Numbers

Ac (III)

Ag (I), (II)

Am (II), (III), (IV), (V), (VI)

Au (I), (III)

Bi (III), (V)

Bk (III), (IV)

Ce (III), (IV)

Cf (III)

Cm (III)

Co (II), (III)

Cr (II), (III), (IV)

Cu (I), (II)

Dy (III)

Er (III)

Es (III)

Eu (II), (III)

Fe (II), (III)

Fm (III)

Gd (III)

Ge (II), (IV)

Hf (IV)

Hg (I), (II)

Ho (III)

In (I), (III)

Ir (I), (II), (III), (IV) (VI)

La (III)

Lu (III)

Md (III)

Mn (II), (III), (IV), (VI), (VII)

Mo (II), (III), (IV), (V), (VI)

Nb (II), (III), (IV), (V)

Nd (III)

Ni (II), (III)

Np (III), (IV), (V), (VI)

No (III)

Os (II), (III), (IV), (VI), (VIII)

Pa (IV), (V)

Pb (II), (IV)

Pd (II), (IV)

Pm (III)

Po (IV)

Pr (III), (IV)

Pt (II), (IV)

Pu (II), (III), (IV), (V), (VI)

Table of Acceptable Stock Numbers
(Continued)

Re (III), (IV), (V), (VI), (VII)

Rh (I), (II), (III), (IV), (VI)

Ru (II), (III), (IV), (V), (VI), (VIII)

Sc (III)

Sm (II), (III)

Sn (II), (IV)

Ta (II), (III), (IV), (V)

Tb (III), (IV)

Tc (IV), (VII)

Th (III), (IV)

Ti (II), (III), (IV)

Tl (I), (III)

Tm (III)

U (III), (IV), (V), (VI)

V (II), (III), (IV), (V), (VI)

W (II), (III), (IV), (V), (VI)

Y (III)

Yb (II), (III)

Zr (IV)

APPENDIX G

SYSTEMS FOR REGISTERING AND NAMING POLYMERS AT CHEMICAL ABSTRACTS SERVICE

by

Robert E. Stobaugh, Warren H. Powell, and Ronald J. Zalac

SYSTEMS FOR REGISTERING AND NAMING POLYMERS AT CHEMICAL ABSTRACTS SERVICE

by

Robert E. Stobaugh, Warren H. Powell, and Ronald J. Zalac

Chemical Abstracts Service, The Ohio State University,
Columbus, Ohio 43210

A computer-based system of registration is being developed for polymers. Through this system polymers will be handled as an integral part of the Chemical Abstracts Service Registry System. The computer-based recognition system will depend on identification of polymer and/or monomer structural units. As is the policy for compounds, the polymers which are identified and indexed in Chemical Abstracts (CA) will be entered into the Registry. CA index nomenclature for polymers has been improved and expanded to keep pace with the Registry System. In addition to improved handling in the CA subject indexes, polymers will be included in CA formula indexes starting with Volume 66 (January-June 1967). The Registry's files of nomenclature (i.e., CA index names, non-CA systematic names, unsystematic names, acronyms, etc.) and bibliographic data link the registered material to the published source documents in which further information can be obtained.

NOT FOR PUBLICATION

SYSTEMS FOR REGISTERING AND NAMING POLYMERS AT CHEMICAL ABSTRACTS SERVICE

Polymers present special problems to those who must organize chemical information. Polymers are unlike most compounds, which usually have fairly compact molecules whose structure can be readily determined and easily characterized in structural diagrams. Instead, polymers normally represent large molecules whose structure is difficult to determine, and therefore is seldom exactly reported. Moreover, a given polymer sample is characteristically a mixture of different structures. This indefiniteness of polymers sets them apart from fully defined compounds and leads to the problems experienced by persons who must design structure-based systems for organizing, naming, and indexing chemical compounds. Since Chemical Abstracts Service is active in all of these areas, we have recently been studying these problems with an eye toward establishing methods for registering polymers and for improving their nomenclature.

Polymer Registration

CAS is now in the process of developing a computer-based Chemical Compound Registry System designed for two purposes (Slide 1): to identify, or recognize, chemical substances on the basis of their structural characteristics, and to file the structural and molecular formulas, names, and bibliographic citations for each substance.

Slide 1

The Registry System is a computer-based identification system which uniquely identifies chemical structures. The Registry Number is assigned to each structure when the structure is first entered into the file. Whenever a structure which is already present on the structure file appears in a new source, the previously assigned number is recovered automatically. This Registry Number functions as a machine address within the associated structure, nomenclature, and bibliographic data files. This information can be correlated with data furnished in the original source through the CA or other reference.

lide 2 The principal file of the Registry System is the Chemical Structure File (Slide 2), a machine-language listing of the structure in terms of the atoms and their connections, or bonds. This record, often referred to as a "connection table," contains all of the information of the two-dimensional projection of the chemical structure. Additional third-dimension information is handled by terms called stereochemical descriptors such as endo and exo, plus and minus. Different stereoisomers are assigned different Registry Numbers. The alternating single and double bonds of cyclic aromatic systems, such as benzene rings in the first two structures, are recognized by computer program as being equivalent, regardless of the particular resonance structure which might be entered.

Routine registration began in 1964 and initially included organic compounds with fully defined two-dimensional structural diagrams, excluding coordination compounds. All such compounds indexed in Chemical Abstracts since 1964 have been registered, as have compounds from several other sources, including reference books and CAS internal files. Some 650,000 different compounds are now on file.

It is the established goal of CAS to extend this Registry System to all substances, including coordination compounds, inorganic compounds, mixtures, and polymers.

As I have noted, the computer recording of polymers presents some interesting problems that are peculiar to this type of chemical substance. Polymers are substances made up of recurring structural units, each of which can be regarded as derived from a specific compound, or monomer. The number of monomeric units is usually large and variable, a given polymer sample being characteristically a mixture of structures with different molecular weights. Thus, polymers are structurally indefinite. I previously described the Registry System as a computer-based identification system which provides a unique identification of chemical structures. However, the word "unique" is equivocal when applied to polymers. Registration of polymers, then, must involve classification into more or less well defined structural families rather than recording of unique polymers.

In order to devise a system for registering polymers within the framework of the established structure-based Registry System, it has been necessary for CAS to determine the type of information on polymers that is routinely available from the literature. Some polymers will be described as fully defined structures, just as most compounds are, while other polymers will have no structural information given. Between these two extremes lie polymers that are described in terms of monomers, or the processes used to make the polymers, or various physical properties, or the significant repeating units. The amount of each type of information available in the literature is important to the design of registration techniques, since it is the goal of registration to provide as much differentiation between substances as possible.

Our earliest studies of polymer literature concentrated on the primary literature. Members of the research staff at CAS found in an analysis of the literature on butadiene polymers that definite structures were given for

23% of the 611 polymers reviewed, 65% were described in monomer terms, and 12% in nonstructural terms (Slide 3). A follow-up survey of entries from trade and commercial compilations, including drugs, foods, feeds, etc., showed that of 2366 polymers found, 21% were described in terms of a significant repeating unit, 15% in terms of the monomers, and 64% in nonstructural terms as shown on Slide 4.

lide 3

lide 4

These studies revealed that many polymers appear in the literature in sufficient detail to permit registration by polymer structure; that is, by using in various combinations the significant repeating unit (SRU) in a polymer chain and the constituent monomers. The significant repeating unit is defined as the smallest group of atoms which on sequential repetition represents the main structure or backbone of the polymer. Registration by nonstructural information is also possible. The following classes of polymers represent the ways in which polymers may be registered:

lide 5 1a) Polymers described by SRU alone, with no monomer information.

Example 1 on Slide 5 shows an SRU without end groups, and Example 2 on Slide 5 shows an SRU with end groups.

lide 6 1b) Polymers described by SRU with monomer information.

See Slide 6.

lide 7 2) Polymers described only in terms of monomers.

See examples in Slide 7.

lide 8 3) Polymers described by non-structural information.

Slide 8 shows polymers described by names, applications, and generic types.

The first class of polymers, SRU's without end groups, will be registered in terms of atoms and connections for machine representation with open-end bonds, i.e., bonds represented, but with no attachments present. Each atom of the structure will be marked as being in a repeated group which has a repeating factor of n. Structures of this type must have two and only two open bonds and a repeating factor of n for acceptance by the computer Edit Program which checks the validity of structural diagrams.

Structure handling of ladder polymers, three-dimensional polymers, and similar types of polymers has not yet been established. That is, studies to develop methods of registering polymers with repeating units which have more than two open bonds, with or without stated end groups, are still in progress.

Since a repeating unit in a polymer may be written in more than one
lide 9 form as shown in Slide 9, the polymer will be treated during computer editing as if the two open bonds were joined to form a ring. This will assure the generation of identical unique connection tables for the same polymer, even though it can be represented by more than one structure drawing.

Polymers described in terms of an SRU with end groups including hydrogen will be input by similar procedures. Each atom of the repeating unit will be marked as being in a repeated group with n as the repeating factor. During the computer edit process the end bonds of the SRU will be treated as if they were joined to form a ring, and the end groups will be treated as disconnected fragments. However, the original attachments to the end groups will be available in the computer record. This method is used to simplify substructure searching in polymers. Cross references will be automatically generated for all members of a set of polymers with the same SRU but different end groups.

A significant repeating unit furnished with monomer information is registered in the same way as that unit without monomer data. For example, ide 10 Slide 10 illustrates poly(ethylene terephthalate) expressed as a significant repeating unit with no monomer information given, as prepared from disodium terephthalate and ethylene dibromide, and prepared from terephthaloyl chloride and ethylene glycol. The unit itself is considered the same and receives the same Registry Number in each case.

Polymers of the second class, for which only monomer information is supplied, receive different Registry Numbers for each constituent monomer, and appropriate cross-references between polymers and monomers will be supplied.

The third type of polymers, those with structural information but described by type or application (Slide 8 repeat), will be registered by name, as assigned by the author. Commonly recognized synonymous names will be given the same Registry Number; otherwise, polymers of this class with different names are considered different and will receive different Registry Numbers.

Criteria for Differentiating Polymers

In developing these techniques for registering polymers, it has been necessary to establish standards as to what is the same and what is different concerning polymers. Polymers made up of structurally isomeric significant
ide 11 repeating units as illustrated in Slide 11 are considered different, and will receive different Registry Numbers. Polymers with the same significant repeating unit and different end groups are considered different; for example,
ide 12 the two structures in Slide 12 will receive different Registry Numbers. And, stereoisomeric forms of polymers when the stereoisomerism is identified in the source document, are considered different and will receive different Registry Numbers.

On the other hand, some polymer characteristics, which are regarded as inconsistent and imprecise for registration purposes, will not be used to differentiate between polymers. These characteristics include physical properties (molecular weight, melting point, viscosity, solubility, color, density, etc.); reaction conditions (pressure, temperature, and solvency); and the ratio of monomers involved in a copolymer (either the ratio of monomers actually incorporated in the polymer or the ratio of charged monomers). Polymers differing only in one or more of these characteristics will be considered the same, and will receive the same Registry Numbers.

"Post-reacted" or "after-treated" polymers will be handled in one of two ways. When the new polymer formed by the post reaction is clearly defined by structure or name, it will be recorded as such and will receive a polymer Registry Number different from that of the original polymer; but, when the new polymer is described in general chemical terms such as oxidized, brominated, etc., or in structural terms, block, graft, cross-linked, etc., it will be recorded as the original polymer and will receive the same polymer Registry Number as that of the original polymer.

lide 8 Slide 8 illustrates this technique. A polymer represented as the significant repeating unit $(-CH_2CH{=}CHCH_2-)_n$ may be stated to be oxidized to form a polymer made up of the unit $(-CH_2COCH_2CH_2O)_n$. In such a case the latter is registered by structure and is different from the original. However, if the unit $(-CH_2CH-CHCH_2-)_n$ is stated simply to be oxidized, the original unit is registered and the oxidized polymer recorded as the original one.

have had very poor acceptance and are not widely used. Other proposals for naming polymers have been made, but none has received the wide acceptance and general usage necessary for a nomenclature system. As a result, most polymers are described by trivial or trade names systems or in terms of the chemical monomer(s) that are used in the preparation of the polymer. Such systems have serious limitations. In most cases, important structural characteristics cannot be described adequately through this type of designation. Furthermore, the same family of polymer compounds may be prepared from a variety of monomers resulting in wide scattering of information in a compilation such as the CA indexes.

For these reasons, the Nomenclature Committee of the Polymer Division of the American Chemical Society has developed a structure-based system for naming polymers that are described in terms of a known, regularly repeating structural unit. When the structure of the repeating unit is unknown, the polymer is described in much the same fashion now used, i.e., by trivial or trade names, or on the basis of the chemical monomers. This system is being evaluated and tested during the preparation of the CA Indexes for Volume 66.

In essence, this structure-based nomenclature system is based on the fundamental structural unit or mer, which is defined as the smallest unit (real or hypothetical) of a polymer. The mer is made up of one or more polyvalent radicals which, when named as polyvalent organic radicals and cited in the directional order specified in the rules, make up the name of the mer. The repeating structural unit of a linear polymer may be composed of a unit that can be expressed in terms of just one polyvalent radical such as methylene or phenylene. Often the mer is more complex and is composed of a series of polyvalent radicals.

The names of linear polymers containing polyvalent radicals are formed by citing the names of the radicals consecutively in a directional specified in the rules. The entire series of polyvalent radicals is prefixed by the term "poly" with suitable enclosing marks. Substituents on the radicals are also denoted by appropriate prefixes. Some examples of names of linear polymers of known structure are shown on Slides 14 and 15.

de 14
de 15

Polymers of unspecified structure are indexed essentially in terms of the monomers. For example, a homopolymer such as poly(1-hexene) whose structure is not described is indexed as "1-hexene, polymer." A copolymer such as a 1-hexene-1-heptene copolymer is indexed as "1-heptene, polymer with 1-hexene" and also as "1-hexene, polymer with 1-heptene."

At present, CAS is conducting development tests of both the registration and naming techniques for polymers. Some polymers are now being registered on an experimental basis by manual procedures that precisely simulate the computer procedures discussed here. These polymers appear in selected sources supplied by the National Library of Medicine and the Food and Drug Administration under Contract C-414 between CAS and the National Science Foundation. Through these procedures, such polymer information as names, Registry Numbers, molecular formulas, and the appropriate source indicator are machine recorded, but structural information is not.

Computer programming for structural information will be undertaken this autumn, and the registration of polymer structures should be possible in the spring of 1968. At that time the polymers selected as entries in the CA indexes for Volume 66 (July-December of 1967) are scheduled to be input to the system. Polymer entries in succeeding CA volume indexes will be registered on a routine basis, as will polymers from the CAS express publications POST-J and POST-P.

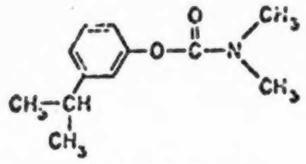
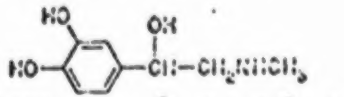
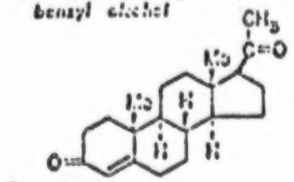
The structure-based nomenclature system for polymers is being evaluated and testing during the preparation of the CA indexes for Volume 66 will also be used, with possible revisions and expansion in succeeding CA volumes.

CHEMICAL COMPOUND REGISTRY SYSTEM

A computer-based system for identifying compounds.
Based on
- a. A Computer-Based Recognition System
- b. Files
  - Structural Data
  - Nomenclature
  - Bibliographic Data

SLIDE 1

CHEMICAL STRUCTURE FILE LISTINGS

COMPOUND STRUCTURE	REGISTRY NO.
 m-Cumenyl dimethylcarbamate	3938-45-2
 3,4-Dihydroxy-α-[(methylamino) methyl]- benzyl alcohol	6912-68-1
 Progesterone	57-83-0

SLIDE 2

RESULTS OF BUTADIENE POLYMER STUDY

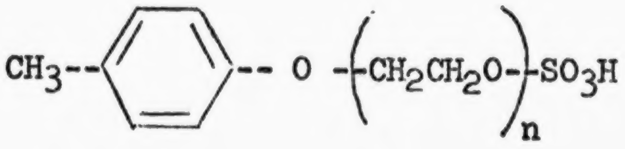
Structure (SRU)	23%
Structure (Monomer)	65%
Nonstructural	12%

RESULTS OF POLYMER SURVEY IN TRADE AND COMMERCIAL COMPILATION

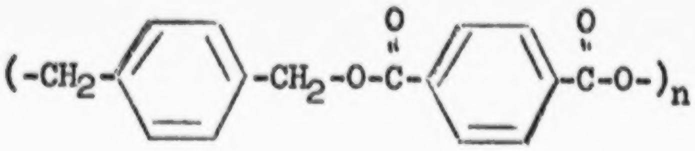
Structure (SRU)	21%
Structure (Monomer)	15%
Nonstructure	64%

POLYMERS DESCRIBED BY SRU ALONE

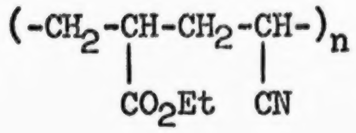
1. With End Groups



2. Without End Groups



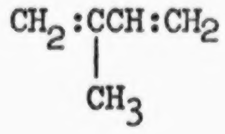
POLYMERS DESCRIBED BY SRU AND MONOMER INFORMATION



$$(\text{-}CH_2\text{-}\underset{\underset{CO_2Et}{|}}{CH}\text{-}CH_2\text{-}\underset{\underset{CN}{|}}{CH}\text{-})_n$$

from $CH_2{:}CH\text{-}CO_2Et$ and $CH_2{:}CH\text{-}CN$

POLYMERS DESCRIBED BY MONOMERS ALONE



1. polymer of $CH_2{:}\underset{\underset{CH_3}{|}}{C}CH{:}CH_2$

2. polymer of $CH_2{:}CHCN$, $PhCH{:}CH_2$, and $CH_2{:}CHCH{:}CH_2$

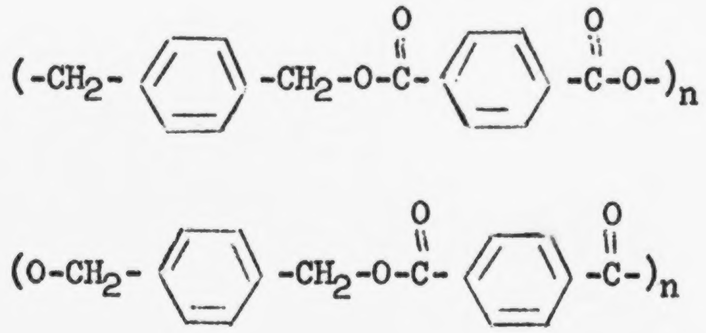
SLIDE 7

POLYMERS DESCRIBED BY NON-STRUCTURAL INFORMATION

1. Deeminite - ion exchange resin
2. Nudek - a plastic
3. Natrimil - a cation exchange resin
4. Cypan - an acrylic polymer
5. Verel - an acrylic fiber

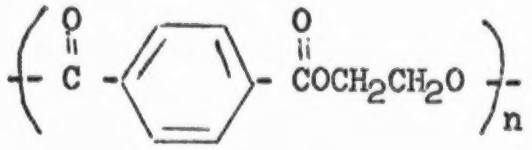
SLIDE 8

ALTERNATIVE FORMS OF POLYMER REPEATING UNIT

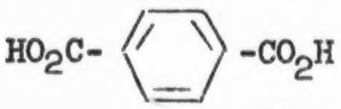

(-CH2-
-CH2-O-C-
O
-C-O-)n
(O-CH2-
-CH2-O-C-
O
-C-)n

DESCRIPTIONS OF POLY(ETHYLENE TEREPHTHALATE)

1. SRU with no monomer information


O
C
O
COCH2CH2O
n

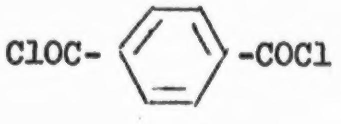
2. SRU with monomer information



and $BrCH_2CH_2Br$

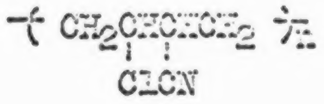
disodium salt

3. SRU with monomer information



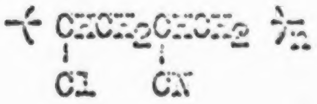
and $HOCH_2CH_2OH$

POLYMERS WITH STRUCTURALLY ISOMERIC SRU's



$-\!\left(CH_2\underset{Cl}{CH}\underset{CN}{CH}CH_2 \right)\!-_n$

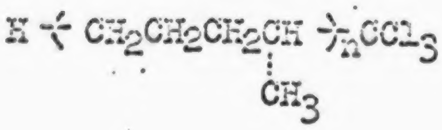
and



$-\!\left(\underset{Cl}{CH}CH_2\underset{CN}{CH}CH_2 \right)\!-_n$

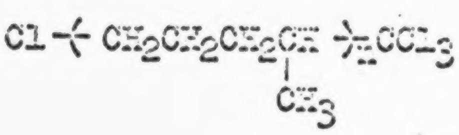
SLIDE 11

POLYMERS WITH THE SAME SRU AND DIFFERENT END GROUPS



$$\mathrm{H}\left(\mathrm{CH_2CH_2CH_2}\underset{\mathrm{CH_3}}{\underset{|}{\mathrm{CH}}} \right)_n \mathrm{CCl_3}$$

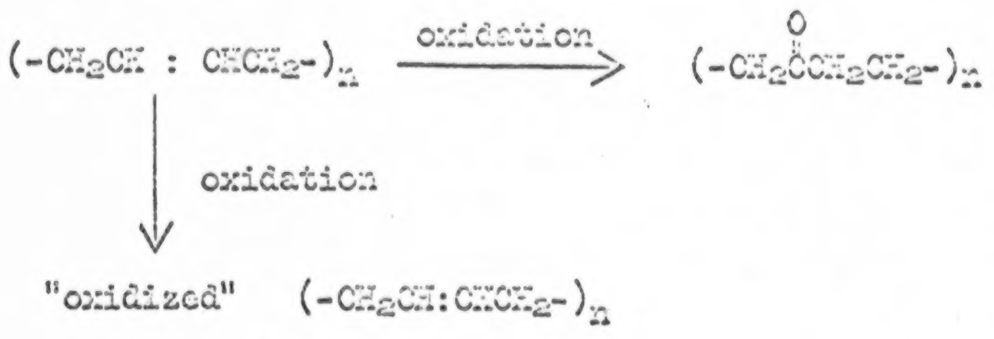
and



$$\mathrm{Cl}\left(\mathrm{CH_2CH_2CH_2}\underset{\mathrm{CH_3}}{\underset{|}{\mathrm{CH}}} \right)_n \mathrm{CCl_3}$$

SLIDE 12

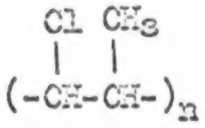
POST-REACTED POLYMER



$$\begin{tikzcd}
(-CH_2CH : CHCH_2-)_n \arrow[r, "\text{oxidation}"] \arrow[d, "\text{oxidation}"] & (-CH_2\overset{O}{\overset{\|}{C}}CH_2CH_2-)_n \\
\text{"oxidized"}\ (-CH_2CH:CHCH_2-)_n &
\end{tikzcd}$$

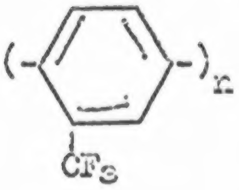
LINEAR POLYMERS COMPOSED OF SIMPLE BIVALENT MERS

$(-CH_2-)_n$ poly(methylene)



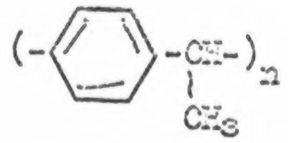
$(-\overset{\mathrm{Cl}}{\mathrm{CH}}-\overset{\mathrm{CH_3}}{\mathrm{CH}}-)_n$

poly(1-chloro-2-methylethylene)

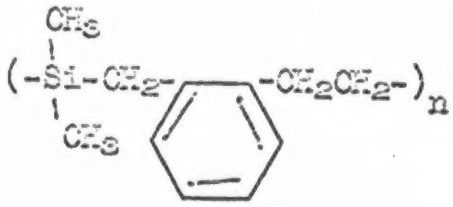


poly[(trifluoromethyl)-*p*-phenylene]

LINEAR POLYMERS COMPOSED OF COMPLEX BIVALENT MERS



poly(p-phenyleneethylidene)



poly[(dimethylsilylene)methylene-o-phenyleneethylene]

SLIDE 15